Development of Tools to Help Understand and Map Protein-Protein Interactions

by

Carlos Henrique Miranda Rodrigues

ORCID: 0000-0002-4420-6401

A thesis submitted in total fulfilment for the
degree of Doctor of Philosophy

in the
Department of Biochemistry and Pharmacology
School of Biomedical Science
THE UNIVERSITY OF MELBOURNE

December 2021

Abstract

Most biological processes are intrinsically coordinated through complex networks of protein-protein interactions. The diversity and the scale of these networks of interactions offer a highly selective and tunable way of modulating protein function. The importance of protein-protein interactions is further highlighted by the fact that many genetic disease-associated mutations are known to be enriched at protein interacting interfaces and, consequently, directly affecting their ability to interact with their partners. My PhD has focused on the systematic analysis of protein-protein interactions as means to provide a deeper molecular understanding of the mechanisms by which they are driven.

I have conducted a large scale analysis of all protein-protein interactions available on the Protein Data Bank and I have shown that, contrary to what was originally thought and corroborating more recent studies, interacting interface regions are not flat and featureless. In fact, a detailed examination of the geometrical and physicochemical properties revealed that the majority of interactions made use of small concavities at the interface, presenting opportunities for single residue sites to be used for fragment targeting, and the development of competitive interface small molecules. In addition, I applied sophisticated data analysis and machine learning techniques to develop novel methods to explore the different mechanisms by which mutations affect protein dynamics, stability and binding affinity to other proteins. These methods have been broadly applied to a variety of different studies, from drug resistance in *Mycobacterium Tuberculosis* to analysing genetic variants in proteins of SARs-CoV-2.

Finally, I investigated the properties of small molecules capable of inhibiting protein-protein complexes. Overall, these were larger than orally available small molecules, R05, and the most potent ones were shown to be enriched with ring substructures, including biphenyls. I explored structure and chemical properties of compounds with inhibitory activity against 23 distinct protein-protein interactions, and used them to propose a novel method to assist in rapid screening and ranking of inhibitors for any protein-protein interaction.

The outcomes of my research provide powerful and valuable biological insights into the nature of protein-protein interactions, improving our understanding of how these can be better targeted for drug discovery and in the context of disease, particularly through mutation characterisation and modulation.

Declaration of Authorship

I, CARLOS HENRIQUE MIRANDA RODRIGUES, declare that this thesis titled, 'DEVELOPMENT OF TOOLS TO HELP UNDERSTAND AND MAP PROTEIN-PROTEIN INTERACTIONS' and the work presented in it are my own. I confirm that:

- the thesis comprises only my original work towards the Doctor of Philosophy except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Carlos Henrique Miranda Rodrigues

December 2021

Preface

I would like to acknowledge the University of Melbourne for funding my research with the Melbourne Research Scholarship. The work in this thesis was in part funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405), the Victorian Government's Operational Infrastructure Support Program, Fundacao de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Cientıfico e Tecnologico (CNPq). I would also like to acknowledge the following people for their contributions towards the thesis:

Chapter 2 consists of the article **Structural Landscapes of PPI Interfaces** which has been submitted for publication to *Proceedings of the National Academy of Sciences of the United States of America* on November 2021. This work was conceived by David Ascher. I was directly involved in the study design, data curation, statistical analysis, methodology and original draft of the manuscript. All authors contributed to manuscript revisions.

Chapter 3 consists of the articles **DynaMut:predicting the impact of mutations on protein conformation, flexibility and stability** published by *Nucleic Acids Research* on July 2018, and **DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations** published in *Protein Science* in September 2020. This work was conceived by David Ascher. I was involved in the study design along with Douglas Pires and David Ascher, and worked on data curation and analysis, investigation, methodology, validation, web server development and wrote the original draft of the manuscript. All authors contributed to manuscript revisions.

Chapter 4 consists of the article **mCSM-PPI2:predicting the effects of mutations on protein–protein interactions** published by *Nucleic Acids Research* on May 2019. David Ascher and Douglas Pires conceived the project. I worked on data curation and analysis, methodology, validation and original draft of the manuscript. I also developed the web server along with Yoochan Myung. All authors contributed to manuscript revisions.

Chapter 5 consists of the article **pdCSM-PPI:using graph-based signatures to identify protein-protein interaction inhibitors** published by the *Journal of Chemical Information and Modeling* on November 2021. David Ascher and Douglas Pires conceived the project. I was involved in the data collection and curation, formal analysis, methodology, validation, web server development, and original draft of the manuscript. All authors contributed to manuscript revisions.

Chapter 6 consists of the article **Kinact: a computational approach for predicting activating missense mutations in protein kinases** published by *Nucleic Acids Research* on July 2018. The project was conceived by Douglas Pires. I was involved in data collection, curation and analysis, as well as methodology design with Douglas Pires and David Ascher. I worked on web server development and original draft of the manuscript. All authors contributed to manuscript revisions.

Chapter 7 consists of the article **mmCSM-PPI: predicting the effects of multiple point mutations on protein-protein interactions** published by *Nucleic Acids Research* on April 2021. This work was conceived by David Ascher. I was involved in the study design along with Douglas Pires and David Ascher. I was responsible for data collection, curation and analysis, methodology, validation, web server development and draft of original manuscript. All authors contributed to manuscript revisions.

Acknowledgements

I would like to thank the University of Melbourne for supporting this PhD research with the Melbourne Research Scholarship and my doctoral supervisors, Associate Professor David Ascher and Associate Professor Douglas Pires, for their constant guidance and incredible support.

David has been a great mentor and a tremendous guide to help me navigate through the world of structural biology. Douglas introduced me to computational and structural biology and his interdisciplinary expertise has been invaluable. Both, David and Douglas, pushed me to "think outside the box" and helped me become a better researcher. I am grateful for all the teachings, discussions, feedback and support they offered me over the past years.

My PhD has brought me great friendships and has given me the opportunity to meet wonderful people from many different places. I am grateful to Malancha Karmakar and Stephanie Portelli for being such amazing friends. These two were without a doubt the best friends one could ever have wished for. I appreciate all the patience they had with me in the beginning and also how much they taught me, not only science, but everything I learned about their respective cultures and food over the years. I would especially like to acknowledge Yoochan Myung for his friendship and assistance with my projects. His hard work has been an inspiration to me. To all my other friends and colleagues - Alex de Sá, Binh Nguyen, Daniella Hock, Elston D'Souza, Michael Silk, Raghad Al-Jarf, Qisheng Pan, Moshe Olshansky and Yara Braga - thank you for all moral support and advice on presentations.

Finally, I would like to thank my family. My parents, Jane Miranda and Carlos Rodrigues, my brother Vinicius Rodrigues and my sister Ana Luísa Rodrigues for their continuous emotional support. It has been difficult to stay such a long period of time far away from them, but they have always been great at making their presence felt via a phone call or video chat.

Publications in Peer Reviewed Journals

1. **Rodrigues, C. H. M.**, Pires, D. E. V., Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic acids research, 46(W1), W350-W355. *Open Access under the Creative Commons Attribution Non-Commercial License*

2. **Rodrigues, C. H. M.**, Ascher, D. B., Pires, D. E. V. (2018). Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic acids research, 46(W1), W127-W132. *Open Access under the Creative Commons Attribution Non-Commercial License*

3. **Rodrigues, C. H. M.**, Myung, Y., Pires, D. E. V., Ascher, D. B. (2019). mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic acids research, 47(W1), W338-W344. *Open Access under the Creative Commons Attribution Non-Commercial License*

4. Karmakar, M., **Rodrigues, C. H. M.**, Holt, K. E., Dunstan, S. J., Denholm, J., Ascher, D. B. (2019). Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PloS one, 14(5), e0217169. *Open Access under the Creative Commons Attribution Non-Commercial License*

5. Pires, D. E. V., **Rodrigues, C. H. M.**, Albanaz, A. T., Karmakar, M., Myung, Y., Xavier, J., ... Ascher, D. B. (2019). Exploring protein supersecondary structure Through Changes in Protein Folding, Stability, and Flexibility. In Protein Supersecondary Structures (pp. 173-185). Humana Press, New York, NY. *By permission of Springer Nature*

6. Myung, Y., **Rodrigues, C. H. M.**, Ascher, D. B., Pires, D. E. V. (2020). mCSM-AB2: guiding rational antibody design using graph-based signatures. Bioinformatics, 36(5), 1453-1459. *By permission of Oxford University Press*

7. Vedithi, S. C., **Rodrigues, C. H. M.**, Portelli, S., Skwark, M. J., Das, M., Ascher, D. B., ... Malhotra, S. (2020). Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of RNA polymerase in Mycobacterium leprae. Computational and structural biotechnology journal, 18, 271-286. *Open Access under the Creative Commons Attribution Non-Commercial License*

8. Karmakar, M., **Rodrigues, C. H. M.**, Horan, K., Denholm, J. T., Ascher, D. B. (2020). Structure guided prediction of Pyrazinamide resistance mutations in

pncA. Scientific reports, 10(1), 1-10. *Open Access under the Creative Commons Attribution Non-Commercial License*

9. Pires, D. E. V., Portelli, S., Rezende, P. M., Veloso, W. N., . . . , **Rodrigues, C. H. M.**, Silk, M. Ascher, D. B. (2020). A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods. In Structural Bioinformatics (pp. 91-106). Humana, New York, NY. *By permission of Springer Nature*

10. Portelli, S., Olshansky, M., **Rodrigues, C. H. M.**, D'Souza, E. N., Myung, Y., Silk, M., ... Ascher, D. B. (2020). Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. Nature Genetics, 52(10), 999-1001. *By permission of Springer Nature*

11. Pires, D. E. V., **Rodrigues, C. H. M.**, Ascher, D. B. (2020). mCSM-membrane: predicting the effects of mutations on transmembrane proteins. Nucleic acids research, 48(W1), W147-W153. *Open Access under the Creative Commons Attribution Non-Commercial License*

12. Pires, D. E. V., Veloso, W. N., Myung, Y., **Rodrigues, C. H. M.**, Silk, M., Rezende, P. M., ... Ascher, D. B. (2020). EasyVS: a user-friendly web-based tool for molecule library selection and structure-based virtual screening. Bioinformatics, 36(14), 4200-4202. *By permission of Oxford University Press*

13. Airey, E., Portelli, S., Xavier, J. S., Myung, Y. C., Silk, M., . . . , **Rodrigues, C. H. M.**, ... Ascher, D. B. (2021). Identifying Genotype–Phenotype Correlations via Integrative Mutation Analysis. In Artificial Neural Networks (pp. 1-32). Humana, New York, NY. *By permission of Springer Nature*

14. **Rodrigues, C. H. M.**, Pires, D. E. V., Ascher, D. B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Science, 30(1), 60-69. *Open Access under the Creative Commons Attribution Non-Commercial License*

15. **Rodrigues, C. H. M.**, D. E. V., Ascher, D. B. (2021). mmCSM-PPI: predicting the effects of multiple point mutations on protein-protein interactions. Nucleic acids research. *Open Access under the Creative Commons Attribution Non-Commercial License*

16. Silk, M., Pires, D. E.V., **Rodrigues, C. H. M.**, D'Souza, E., Olshansky, M., Thorne, N. Ascher, B. D. (2021). MTR3D: Identifying regions within protein tertiary structures under purifying selection. Nucleic Acids Research. *Open Access under the Creative Commons Attribution Non-Commercial License*

17. Potter, E. L., **Rodrigues, C. H. M.**, Ascher, D. B., Abhayaratna, W. P., Sengupta, P. P., Marwick, T. H. (2021) Machine Learning of ECG Waveforms to Improve Selection for Testing for Asymptomatic Left Ventricular Dysfunction. JACC: Cardiovascular Imaging. *By permission of Elsevier*

18. **Rodrigues, C. H. M.**, Pires, D. E. V., Ascher, D. B. (2021). pdCSM-PPI: using graph-based signatures to identify protein-protein interaction inhibitors. Journal of Chemical Information and Modeling. *By permission of American Chemical Society*

Articles Submitted for Publication

1. **Rodrigues, C. H. M.**, Pires, D. E. V., Ascher, D. B. (2021). Structural Landscapes of Protein-Protein Interaction Interfaces. Submitted to *PNAS* on November 2021

2. Karmakar, M., Cicaloni, V., **Rodrigues, C. H. M.**, Spiga, O., Santucci, A. David B. Ascher. (2021). HGDiscovery: an online tool providing functional and phenotypic information on novel variants of homogentisate 1,2-dioxygenase. Submitted for publication to *Briefings in Bioinformatics* on April 2021

Contents

List of Figures

List of Tables

Abbreviations

Ala	**Ala**nine
Arg	**Arg**inine
AUC	**A**rea **U**nder the **C**urve
BLI	**B**io-**L**ayer **I**nterferometry
CV	**C**ross-**V**alidation
FN	**F**alse **N**egatives
FP	**F**alse **P**ositives
GP	**G**aussian **P**rocess
IG	**I**nformation **G**ain
ITC	**I**sothermal **T**itration **C**alorimetry
KB	**K**nowledge **B**ased
MCC	**M**atthews **C**orrelation **C**oefficient
ML	**M**achine **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
NA	**N**ot **A**available
NMA	**N**ormal **M**ode **A**Analysis
NMR	**N**uclear **M**agnetic **R**esonance
nsSNV	**n**on-**s**ynonymous **S**ingle **N**ucleotide **V**ariant
PCA	**P**rincipal **C**omponent **A**nalysis
PDB	**P**rotein **D**ata **B**ank
PPI	**P**rotein-**P**rotein **I**nteractions
PPIN	**P**rotein-**P**rotein **I**nteraction **N**etwork
RFE	**R**ecursive **F**eature **E**limination
RMSD	**R**oot **M**ean **S**quared **D**eviation

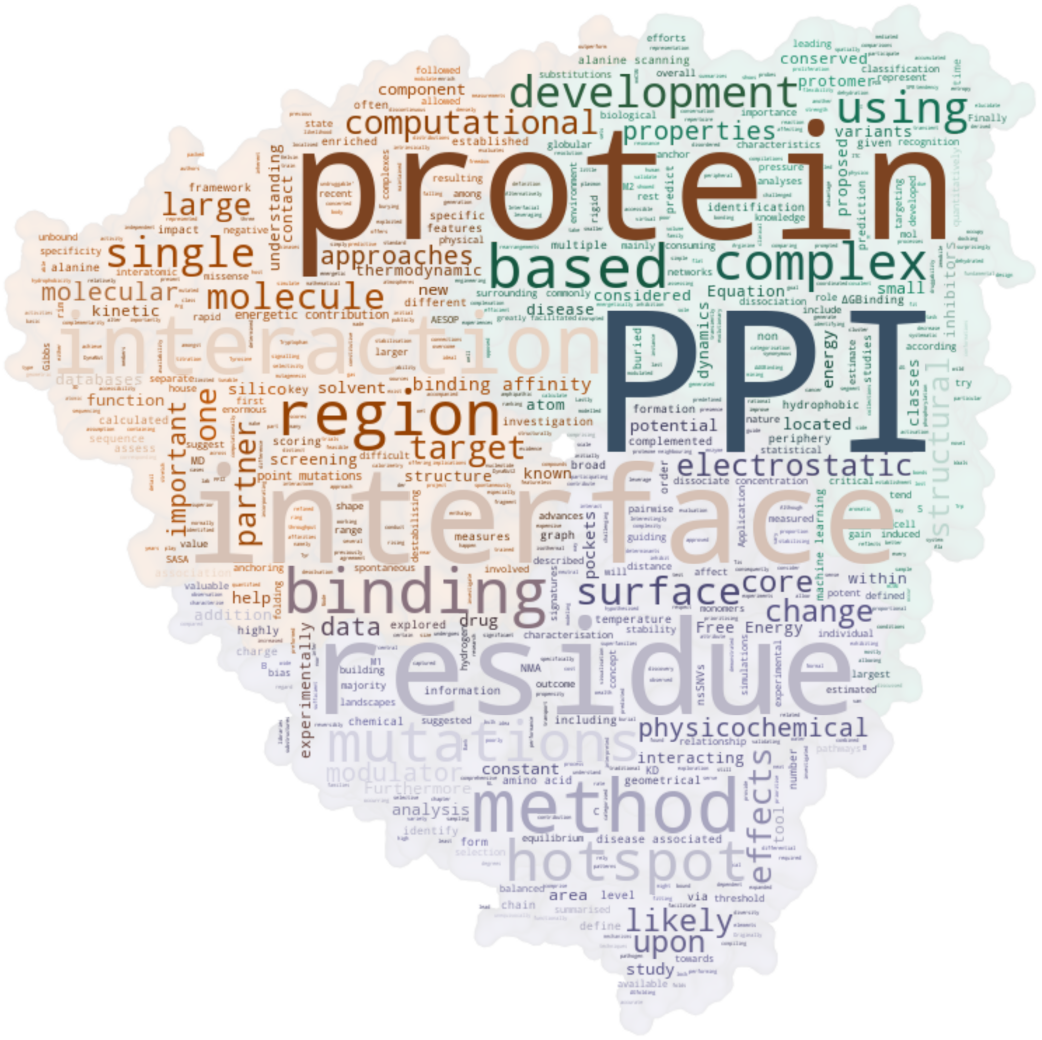
RMSE	**R**oot **M**ean **S**quared **E**rror
RO5	Lipinski **R**ule **o**f **5**
SASA	**S**olvent **A**ccessible **S**urface **A**rea
SPR	**S**urface **P**lasmon **R**essonance
SVM	**S**upport **V**ector **M**achines
TN	**T**rue **N**egatives
TP	**T**rue **P**ositives
TPR	**T**rue **P**ositive **R**ate
FPR	**F**alse **P**ositive **R**ate
Trp	**Tryp**tophan
Tyr	**Tyr**osine

Symbols

Å	Angstrom
ρ	Pearson's Coefficient Correlation
σ	Standard Deviation

Chapter 1

LITERATURE REVIEW



1.1 Protein-Protein Interactions

1.1.1 Introduction

Proteins are involved in most fundamental biological processes, including cell proliferation [1], signalling [2], host-pathogen interactions [3] and protein transport [4], which are all intrinsically coordinated through complex networks of protein-protein interactions (PPIs).

Each protein will often interact through specific regions with several different protein partners (Figure 1.1). While the size of the human proteome is estimated to be ≈20,000, the number of interactions, also known as the interactome, has been estimated to be over 650,000 [5], offering a highly selective and tunable way to modulate protein activities and pathways.

1.1.2 PPI Interfaces

The physical association of two or more protein surfaces forms a protein-protein interface, which includes the area of surfaces between protomers that are buried from solvent (the interface core), and the regions of each protomer surrounding the core that participate in non-bonding interactions with partner proteins (the interface periphery or rim region). Interfacial residues are more conserved than the rest of the protein surface, and, more specifically, residues located in interface core regions (those buried from bulk solvent) tend to be more conserved than those in the periphery surrounding the core [7]. This differential rate of conservation of surface residues has been used to predict likely protein interaction interfaces [8]. Furthermore, PPI interfaces have been described as exhibiting electrostatic charge complementarity [9], and also being 80-90% dehydrated with water molecules participating in interactions mostly with residues at the interface periphery, as opposed to near 100% dehydration in the hydrophobic cores of globular protein monomers [7].

Originally, interface regions were considered to be large (400 $Å^2$ to 4000 $Å^2$), hydrophobic, flat and featureless [10], leading to their classification as poor targets for the development of small molecule modulators. This idea, however, has been challenged through

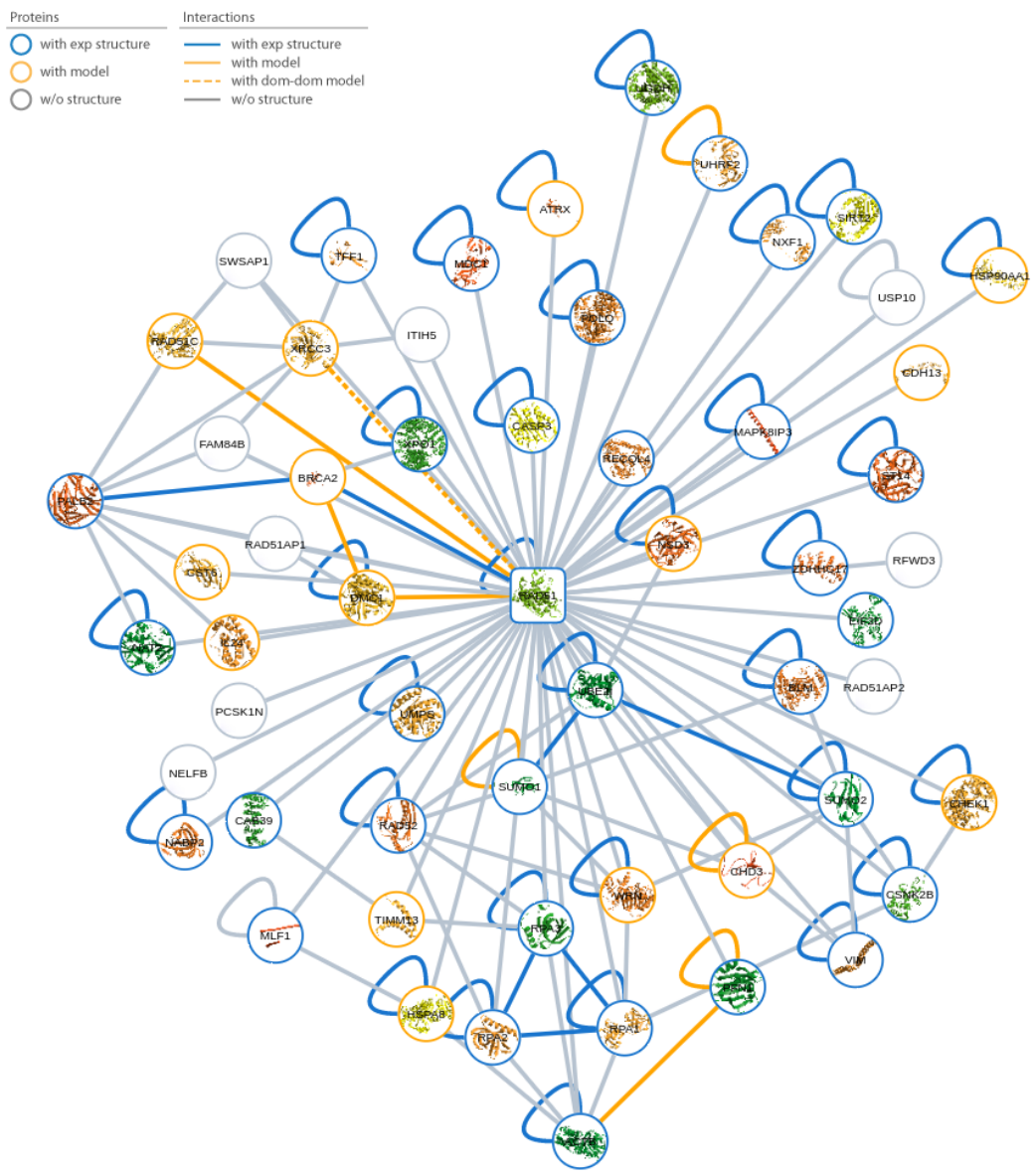


FIGURE 1.1: Interactome of Homo Sapiens DNA repair protein RAD51 homolog 1. A single protein can interact with many other protein partners to coordinate different functions in different signalling pathways. Only first-neighbour interacting partners of RAD51 (highlighted in red in center) are shown. Node borders and edges in the network are coloured by whether or not the protein or interaction respectively have experimental or computationally modelled structural data. Protein structures, when available, are shown as a cartoon representation within the node. The mix of interactions with and without structures is a product of the lack of comprehensive structural coverage for PPI networks. Generated using Interactome3D [6].

their structural and thermodynamic characterisation [11], and has also allowed the classification of these regions into three broad classes of interactions: (A) interaction between globular proteins, relatively rigid, interacting with little or no change to their structure; (B) at least one of the preformed proteins undergoes a significant change in structure

upon binding (induced fit); and (C) concerted folding upon binding, a disordered region of a protein folds upon interaction with another partner (Figure 1.2). The interfaces of interactions falling into the first two classes have been shown to comprise binding regions located in more than one discontinuous segment of the protein sequence, while those of the last class are normally made through binding regions located on a single sequence stretch. In addition, the structural characterisation of PPI interactions allowed for further categorisation of pairwise interactions according to the nature of interacting partners and interface residues as summarised in Figure 1.3.

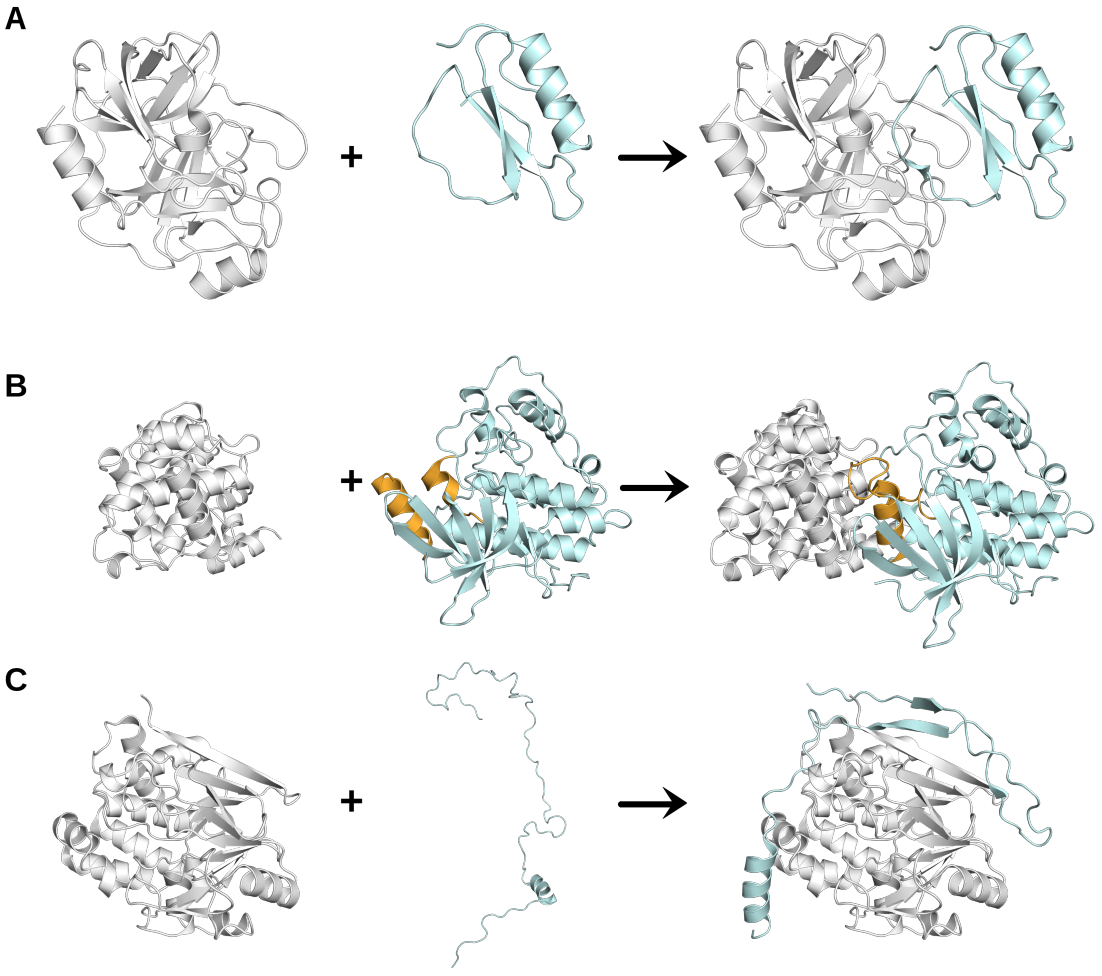


FIGURE 1.2: Types of PPIs according to structural and thermodynamic characterisation. The first group (A) represents association between preformed structures, relatively rigid, that suffer none or little conformational changes upon complexation (rBTI-trypsin complex, PDB IDs: 2A7H, 3RDY and 3RDZ). The second group (B) depicts interactions in which at least one of the partners undergoes under conformational changes (highlighted in yellow) (CDK2 cyclin A in complex with a substrate peptide, PDB IDs: 1VIN, 2R3I and 2CCH). Finally, the third type (C) occurs when a disordered region folds upon interaction with another protein (Phosphatase 1 alpha (PP1) in complex with Spinophilin, PDB IDs: 2FFT and 3EGG).

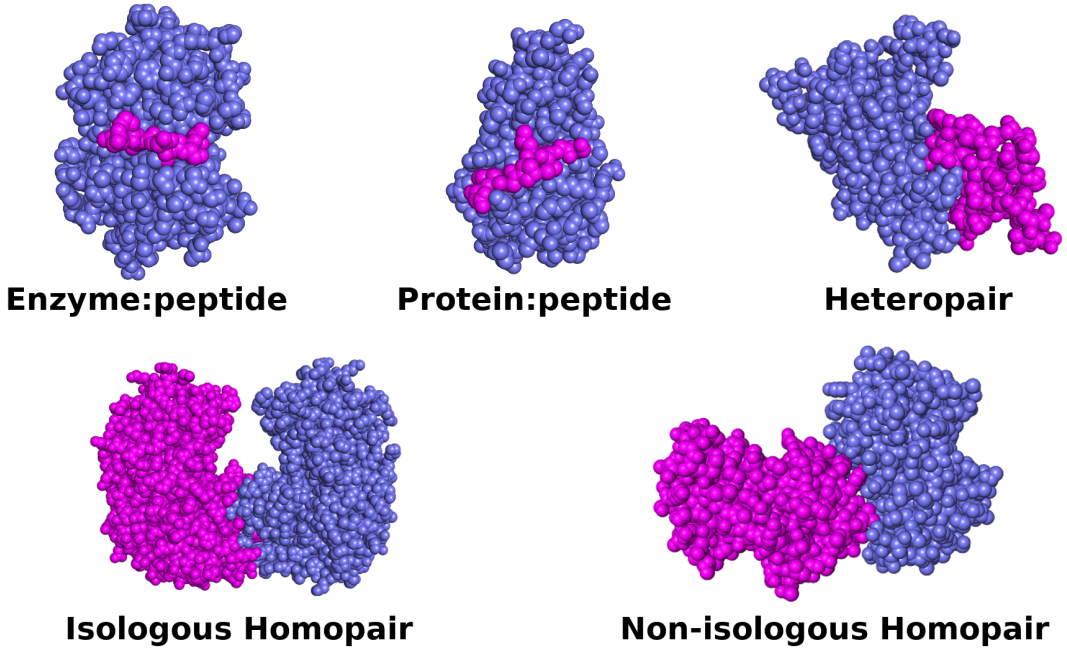


FIGURE 1.3: Interface types according to structural characterisation of pairwise PPIs. Isologous homopairs (PDB ID: 3FPC) represent associations of two proteins with the same sequence (or highly identical sequences - 95% identity), using nearly identical residues on each side of the interface. Non-isologous homopairs (PDB ID: 1CZY) comprise interactions between two proteins with high sequence similarity, but using different interface residues. Association of two distinct globular proteins were classified as heteropairs (PDB ID: 1OQD). Interfaces the interacting pair comprised a globular protein and a peptide (sequence length <41) were further split into Enzyme:peptides (PDB ID: 8PCH) and Protein:peptides interactions (PDB ID: 1P4U).

The ability to identify residues that form the interface of PPIs is critical for understanding the structural and physicochemical determinants of protein recognition and binding affinity. It also has wide applications in modeling and validating protein interactions predicted by high-throughput methods, in engineering proteins, and in prioritising drug targets [12–14].

Protein complex interfaces can be defined using two different *in-silico* methods. The first simply defines residues as being at the interacting interface by the physical distance between these and any other residue from the partner protein, using a threshold value which can vary according to the study, but it is commonly set to 5 Å[15–17]. Alternatively, one can use the assumption that interface regions are buried from solvent, allowing for them to be identified by comparing the difference in solvent accessible surface area ($SASA$) at the residue level between the individual protomers and within the complex (Equation 1.1) [18]. Residues that show a decrease superior to a predefined

threshold, commonly set between 4-5% [19, 20], in their $SASA$ upon complexation are then considered to be part of the interface.

$$SASA_{AB} = SASA_A + SASA_B - SASA_{AB} \tag{1.1}$$

A and B are protein chains in a pairwise interaction.

At the molecular level, PPIs are mainly established through electrostatic contacts by the atoms on the side chains of the amino acid residues comprising each protein interface [21]. Figure 1.4 shows in detail two monomers (single chain proteins) and the interatomic interactions among the residues located on their region of binding. Over 75% of contacts are formed based on side-chain structures of the residues at the interface.

1.1.3 Thermodynamics and Kinetics of Protein-Protein Binding

The strength of PPIs is mainly dictated by the repertoire of molecular interactions established and lost in complex formation [23]. In the study of the kinetics of molecular complexes, the binding affinity between two molecules, M_1 and M_2, can be experimentally quantified by the dissociation constant (K_D), which measures the propensity of a larger complex to separate (dissociate) reversibly into smaller components (Equation 1.2).

$$K_D = \frac{[M_1] \times [M_2]}{[M_1 M_2]} \tag{1.2}$$

where $[M_1]$, $[M_2]$ and $[M_1 M_2]$ represent the equilibrium concentrations of the two molecules (M_1 and M_2) and the complex ($M_1 M_2$), respectively.

Molecular affinities can also be experimentally determined by Isothermal Titration Calorimetry (ITC), from Surface Plasmon Resonance (SPR) kinetic equilibrium constants, or by analysing interference patterns of white light reflected from the surface of a biosensor tip, namely Bio-Layer Interferometry (BLI). These can be quantitatively calculated using K_D by the Gibbs Free Energy of binding ($\Delta G^{Binding}$) under standard conditions (1.0 *mol* initial concentration of bound and unbound complex and its components, 1

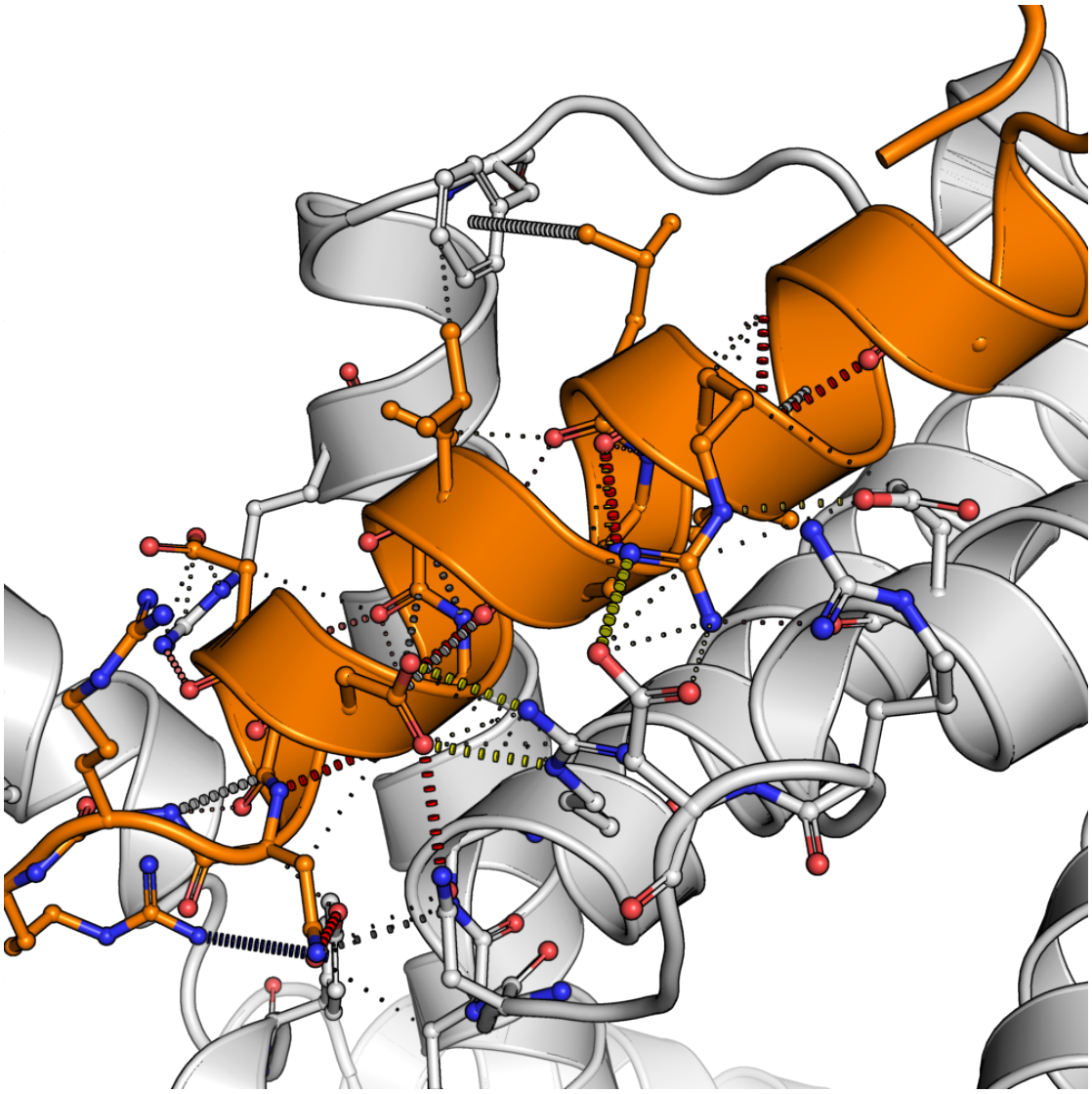


FIGURE 1.4: PPI complex of Bcl2 and Bak proteins. Both proteins (Bcl-xL and Bak BH3 complex, PDB: 5FMJ) are depicted using cartoon representation with Bcl2 and Bak coloured in grey and orange, respectively. Different types of interatomic interactions are illustrated by different colours. Ionic interactions are shown in yellow, red for hydrogen bonds and blue for halogen bonding interactions. The thickness of each dash is proportional to the distance of the between interacting atoms. Other types of interactions were hidden for visual purposes. Interactions calculated using Arpeggio [22]

atmosphere of pressure, and at a temperature of 298.15 *Kelvin*). The mathematical equation that defines $\Delta G^{Binding}$ is defined is represented in Equation 1.3.

$$\Delta G^{Binding} = -R \times T \times \ln K_D \tag{1.3}$$

R is the ideal gas constant $\approx 1.9872\text{cal} \times K^{-1} mol^{-1}$, T is the temperature in *Kelvin*, and K_D is the equilibrium dissociation constant, which measures the potential of a

complex of two proteins to dissociate into its separate components (Equation 1.2) as a spontaneous reaction.

In addition, the formation of molecular interactions can also be interpreted in terms of entropy (S), which relates to the degrees of freedom for the rearrangements of elements, being a log measure of the possible configurations of the system; and enthalpy (H), representing the sum of the internal energy of the system and is related to the establishment of van der Waals', hydrogen and charge interactions [24]. These two concepts can be modelled as a thermodynamic state function that quantitatively measures the likelihood of association between two molecules spontaneously occurring at constant pressure and temperature, also known as Gibbs Free Energy of folding(ΔG) (Equation 1.4). In this sense, associations resulting in negative ΔG values are spontaneous and the more negative the more likely to happen.

$$\Delta G = \Delta H - T \times \Delta S \tag{1.4}$$

ΔH represents the change in enthalpy on binding in $Jmol^{-1}$, ΔS depicts the change in entropy on binding in $JK^{-1}mol^{-1}$ and T is the temperature in Kelvin.

1.1.4 Hotspots

An important observation with respect to the energetics and kinetics of PPI binding was that key interface residues, namely "hotspots", contribute to a large proportion of the overall binding energy. Experimentally, the identification of hotspots is measured via a process known as Alanine Scanning [25], all wild-type interface residues are mutated to alanine (Ala), one at the time, followed by measurements of changes in Binding Free Energy ($\Delta G^{Binding}$) [26] (Figure 1.5). Because Alanine has a small side chain, substitutions to this amino-acid are likely to affect interatomic interactions and consequently impact on the energetic contribution of a particular residue.

Hotspots residues are often large, amphipathic residues, such as Tyrosine (Tyr), Tryptophan (Trp) and Arginine (Arg), that form central connections in structurally conserved, densely-packed residue interaction networks at the PPI interface [28, 29]. However,

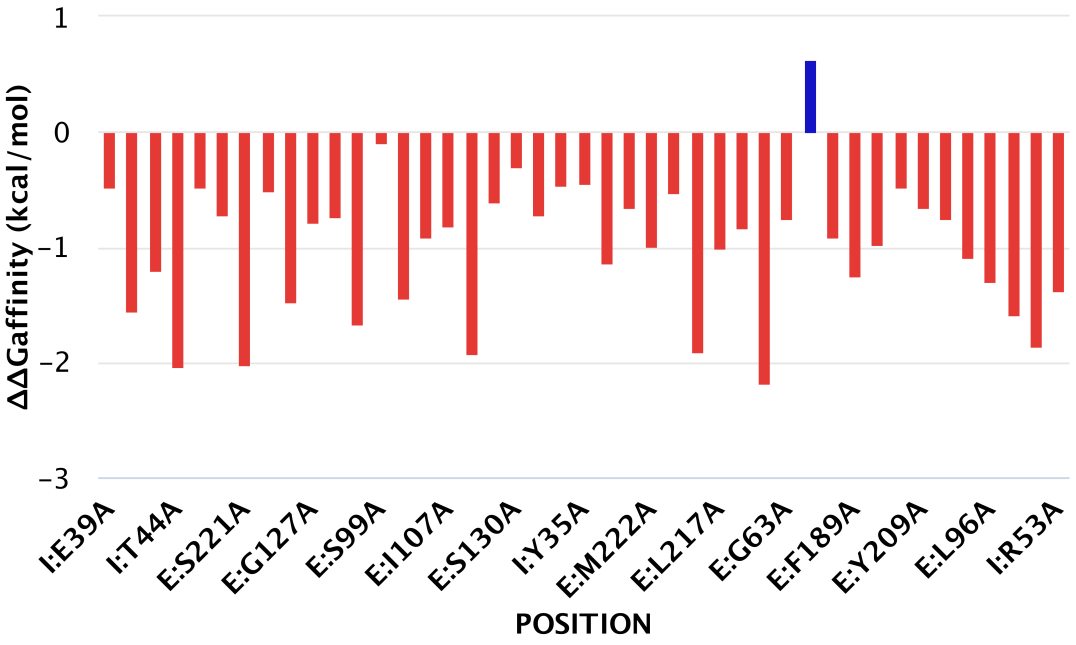


FIGURE 1.5: Summary of Alanine Scanning results for the PPI complex of Subtilisin Carlsberg and Eglin C proteins. Most substitutions to Ala decrease binding affinity with negative changes in Gibbs Free Energy of Binding ($\Delta G^{Binding}$). Positions mutations have mild effects have been hidden in the x-axis for visualisation purposes. Generated using mCSM-PPI2 [27].

no single attribute (shape, charge or hydrophobicity, for instance) has been found to unequivocally define a hotspot [30, 31].

The computational and experimental exploration of interface areas has refined the traditional definition of hotspots, with PPI interface landscapes containing important "complemented pockets" and "anchor" residues (Figure 1.6). Complemented pockets are defined as small pockets on PPI surface with sufficient volume to occupy a single residue of a partner protein [32]. Anchor residues are residues burying the largest proportional of their surface area upon binding, and are hypothesised to be critical in rapid "lock and key" complex formation, followed by induced fitting of the rest of the interface [33]. These residues often play a very important role in both specificity and binding affinity.

1.1.5 Disease Mutations on PPIs

Studies have shown that disease-associated non-synonymous single nucleotide variants (nsSNVs) are enriched in both protein cores and protein interaction interfaces [34–36].

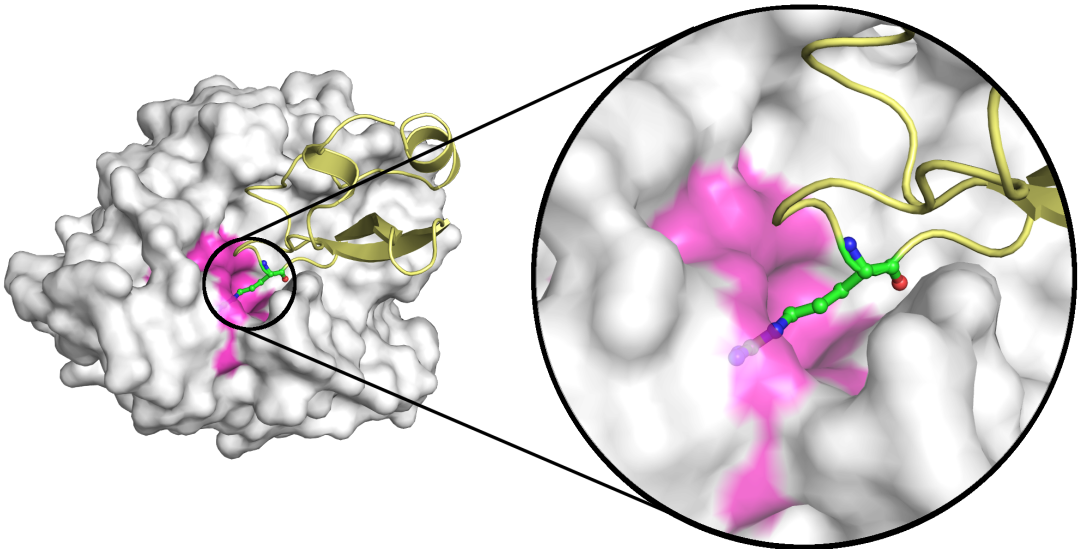


FIGURE 1.6: Complemented pocket at the interacting interface of the complex of blood Coagulation Factor VIIa (grey) and mutant BPTI-5L15 (yellow) (PDB: 1FAK). Protruding Arg15 (green) from BPTI-5L15 occupies the complemented pocket (magenta) at the interface region of the blood coagulation factor VIIa.

Interestingly, a study with 100.000 disease-associated variants, has suggested that while the majority of these mutations can alter PPIs, affecting binding affinity or resulting in novel interactions, they maintained protomer stability [37]. Furthermore, neutral nsSNVs were observed to be more likely localised at peripheral/rim regions. Not surprisingly, hotspot residues, which provide the largest energetic contribution towards binding, were enriched amongst disease-associated nsSNVs [38].

The importance of understanding the relationship between disease mutations and PPIs accompanied by advances in next-generation sequencing techniques has prompted the development of comprehensive databases of experimentally measured binding energy changes, most of which derived from alanine scanning mutagenesis, and thermodynamics properties of PPIs. These large compilations of data represent valuable sources for mining information that can be used to help assess individual energetic contribution of residues at PPI interfaces, and more importantly help elucidate the mechanisms by which missense variants affect PPIs leading to diseases. Table 1.1 summarises a selection of databases compiling information on the effects of mutations on PPIs.

Name	**Description**	**Reference**
SKEMPI2	It comprises a total of 6,187 unique entries (single and multiple mutations), of which more than half corresponded to mutations to alanine	[39]
SKEMPI	Database of experimental measurements of changes in $\Delta G^{Binding}$ upon single variants	[40]
PROXiMATE	Database of effects of missense mutations in heterodimeric PPIs with structural and functional information	[41]
ABbind	Database of antibody-antigen, antibody-effector and antibody-like protein complexes, including non-alanine mutations	[42]
dbMPIKT	Database of experimental kinetic and thermodynamics data on of mutant protein interactions	[43]
ASEdb	Manually curated database of experimentally derived hotspots (from alanine scanning mutagenesis studies)	[44]
PCRPi	Database of computationally predicted hotspots	[45]
ProTherm	Database of experimentally determined thermodynamic parameters of protein stability changes upon mutations	[46]
ThermoMutDB	Manually curated database of thermodynamic and kinetic parameters for wild-type and mutant proteins	[47]

TABLE 1.1: Summary of comprehensive databases compiling experimental data on the effects of missense mutations on PPIs

1.1.6 Predicting the Effects of Mutations on PPIs

Given the importance of understanding the role of mutations on PPIs and the time-consuming task of assessing these effects with wet-lab experiments, a number of computational tools that try to predict the effects of mutations *in-silico* have also been proposed. The development and evaluation of these approaches has been greatly facilitated by the rising of large collections of data, such as the ones described previously. These computational approaches can be categorised into 4 broad classes: molecular dynamics simulations, energy-based, machine learning-based and other methods.

Molecular Dynamics (MD) has been used to simulate alanine substitutions and estimate the corresponding changes in $\Delta G^{Binding}$ [48]. This approach has been used to suggest that energetically important residues are highly immobile and tend to cluster in functionally important regions involved in protein-protein recognition [49]. However,

the enormous computational cost of MD simulations led the authors to suggest that it should not be used as a sole method for hotspots identification.

Energy-based methods rely upon a free energy scoring function to estimate the energetic contribution to binding for every interface residue [50, 51]. In addition, a study using rigid-body docking with electrostatics and desolvation scorings have also been used to generate energy scoring functions, which calculate the tendency of a given residue to be located at the interface and infer potential hotspots [52].

Machine learning-based (ML) approaches leverage the wealth of information on thermodynamic and kinetic characteristics of PPIs accumulated over the years, and try to identify complex relationships that can then be used to characterise hotspots and predict the effects of mutations. These are known for using a broad range of distinct features, including physicochemical properties and shape specificity [53], solvent accessibility and statistical potentials [54], distance patterns among residues from graph-based representation of residue environment [55], and atom contacts, atom contact areas from the target residue and its neighbouring residues [56].

Lastly, a variety of other methods have also been proposed. These range from more basic approaches, which only consider simple geometric characteristics of interface residues to investigate changes in burial level of atoms from interface before and after binding [57], to investigation of spatially conserved physico-chemical interactions (hydrogen bonds, hydrophobic and aromatic interactions) within PPI families [58], and knowledge-based (KB) methods [59]. Finally, a more recent computational framework, AESOP (Analysis of Electrostatic Structures of Proteins) [60], has also been proposed in order to allow *in-silico* alanine scanning and facilitate the investigation of electrostatics in PPI. AESOP evaluates the contribution of single amino acids based on electrostatic potential functions generated from family-based comparisons.

In terms of performance, the outcome for these methods are highly dependent on the data in which they were trained, performing poorly on independent test sets. In addition, studies have shown that in most cases these present a certain bias towards destabilising mutations, given the inherent bias within the databases used for building such methods [61]. Finally, many diseases, like cancer, might occur due the presence of multiple mutations [62, 63], and the majority of approaches available for *in-silico* predictions are

limited to analysis of single point mutations. A selection of methods to assess the impact of mutations on PPIs is summarised in Table 1.2 and 1.3.

Name	**Mutation type**	**Data set**	**Approach**	**Correlation**	**Reference**
mCSM	Single	ProTherm	ML	0.67	[64]
DUET	Single	ProTherm	ML	0.71	[65]
SDM	Single	ProTherm	KB	0.61	[59, 66]
SAAFEC-SEQ	Single	ProTherm	KB+MD	0.61	[67]
SAAFEC	Single	ProTherm	KB+MD	0.65	[68]
ENCoM	Single	*NA*	NMA	*NA*	[69]
I-Mutant	Single	ProTherm	ML	0.71	[70]
FoldX	Single and Multiple	ProTherm	Potential energy function	0.73/0.80	[71]

TABLE 1.2: Computational approaches to assess the effects of mutations on $\Delta\Delta G^{Folding}$

Name	**Mutation type**	**Data set**	**Approach**	**Correlation**	**Reference**
mCSM	Single	SKEMPI	ML	0.80	[64]
iSEE	Single	SKEMPI	ML	0.80	[72]
AESOP	Single	*NA*	Scoring function	*NA*	[60]
BeAtMuSiC	Single	SKEMPI	Statistical potentials	0.40	[73]
MutaBind	Single	SKEMPI	Scoring function	0.77	[74]
Rosetta	Single	Alanine Scanning (230)	Scoring function	0.75	[50]
MMPBSA	Single and Multiple	*NA*	MD	*NA*	[75]
BindProfX	Single and Multiple	SKEMPI	Scoring function	0.73/0.69	[76]
MutaBind2	Single and Multiple	SKEMPI	Scoring function	0.76/0.74	[77]
FoldX	Single and Multiple	SKEMPI	Energy function	0.81/0.37	[71]
ZEMu	Single and Multiple	SKEMPI	MD	0.53/0.65	[78]

TABLE 1.3: Computational approaches to assess the effects of mutations on $\Delta\Delta G^{Binding}$

1.1.7 PPIs as Drug Targets

The enormous diversity of PPIs within the cells offers potential for the development of chemical and biological modulators that target specific pathways, through either the inhibition or stabilisation of specific interactions [79]. Although PPIs were initially considered 'undruggable' targets, a number of PPI inhibitors are now approved or in clinical trials [80], especially through targeting hotspots at the interacting interface [11]. These take advantage of the selectivity that can be difficult to achieve through inhibitors of members of enzyme superfamilies.

Application of new drug development approaches (fragment-based screening, targeting hotspots and increased structural data availability) has greatly facilitated PPI modulator development efforts. Furthermore, efforts to enrich screening libraries that can serve as probes for guiding rational design of new and more potent PPI modulators have been proposed, mainly via statistical analyses of physicochemical properties of compounds with activity over PPIs [81, 82], and the development of publicly available databases of PPI small molecules modulators [83]. However, PPIs are still considered very challenging drug development targets, and advances in new methods and computational tools can help guide and improve PPI modulator development.

In order to overcome the complexity of using PPIs as drug targets, the concept of complemented anchoring pockets at interface regions has been exploited for research into PPI drug discovery [84–86]. Furthermore, a study has suggested that such anchoring pockets may only exist transiently in the unbound protein, which make them difficult to be captured in virtual or experimental chemical screening [87]. In this regard, computationally expensive MD methods are required to sample transient states at atomic resolution, which is time consuming and not feasible for large scale analysis [88].

1.2 Aims

The overall goal of this project is to conduct a systematic analysis of the physicochemical and geometrical nature of PPI interfaces, including through the development of new computational methods, in order to gain a better understanding of how PPIs are disrupted in disease and can be modulated.

In Chapter 2, I explored the structural landscapes of PPI interfaces based on the PDB. I calculated geometrical and physicochemical properties of PPI interfaces using *in-silico* tools and compared the distributions of these features across different classes of interfaces, and discussed implications for druggability.

In Chapter 3, I developed two methods, DynaMut and DynaMut2, to help understand the effects of missense mutations on protein stability and flexibility. By incorporating the dynamics component, using Normal Mode Analysis (NMA), these methods have shown more balanced predictions for stabilising and destabilising mutations. Furthermore, both methods include analyses and visualisation of protein dynamics by sampling conformations via NMA using protein 3D structure.

In Chapter 4, I explored the effects of single-point mutations on PPI binding affinity, working from experiences using in-house framework of graph-based signatures. Here I demonstrated how my method, mCSM-PPI2, was able to outperform more than 23 other methods and was also shown to be valuable in identifying hotspots at PPI interfaces.

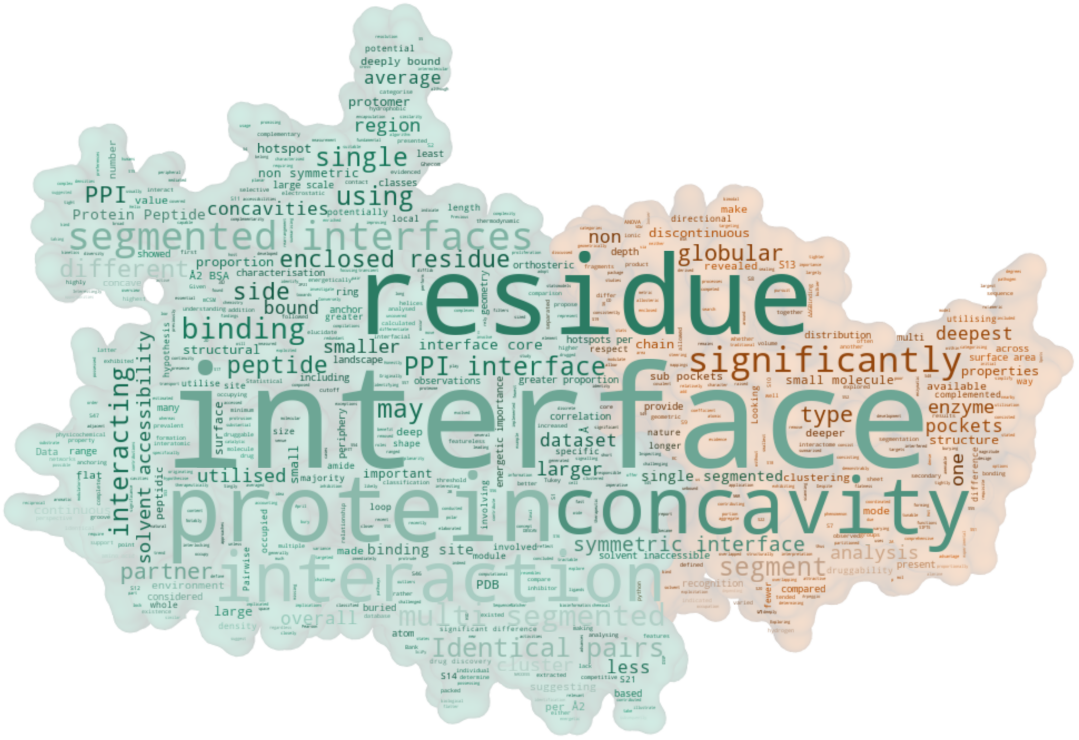
In Chapter 5, I expanded the in-house framework of graph-based signatures to rapid identification and ranking of small-molecules likely to inhibit PPI complexes. In agreement with previous studies and building on more recent data, I showed that more potent PPI inhibitors are larger and enriched with complex ring substructures. The outcome of this chapter will be an important tool for guiding more efficient screening of new PPI inhibitors.

In Chapter 6, I investigated the effects of single-point mutations on phosphorylation-mediated interactions. Here I explored structural- and sequence-based properties of protein kinases to help identify gain of function mutations likely to lead their constitutive activation, as well as prioritise variants for further investigation.

In Chapter 7, leveraging knowledge from Chapters 3 and 4, I developed a predictive method to assess the impact of multiple-point mutations on PPI binding affinity. I calculated physicochemical and geometrical properties for multiple residue environments and combined them with evolutionary scores, dynamics features from NMA and non-covalent contacts, as evidence to train and validate a machine learning method, mmCSM-PPI, which has shown to have a balanced and accurate prediction.

Chapter 2

Structural Landscapes of Protein-protein Interaction Interfaces



PNAS
www.pnas.org

Main Manuscript for

Structural Landscapes of PPI Interfaces

Carlos H. M. Rodrigues [1,2,3], Douglas E. V. Pires [1,2,4], Tom L. Blundell [5], David B. Ascher [1,2,3,5,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria
[2] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria
[3] School of Chemistry and Molecular Biosciences, Bio21 Institute, University of Queensland, Brisbane, Victoria
[4] School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria
[5] Department of Biochemistry, University of Cambridge, Cambridge, UK

[*]To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au.

Author Contributions: D.B.A. designed research; C.H.M.R. and D.B.A. performed research; C.H.M.R. analysed data; and C.H.M.R., D.E.V.P., T.L.B. and D.B.A. wrote the paper.

Competing Interest Statement: The authors declare no competing interest.

Classification: Biological Sciences, Biochemistry

Keywords: protein-protein interface, structural biology, protein binding site, drug design

This PDF file includes:

Abstract

Proteins are capable of highly specific interactions and are responsible for a wide range of functions, making them attractive in the pursuit of new therapeutic options. Previous studies focusing on overall geometry of protein-protein interfaces, however, concluded that PPI interfaces were generally flat. More recently this idea has been challenged by their structural and thermodynamic characterisation, suggesting the existence of concave binding sites that are closer in character to traditional small-molecule binding sites, rather than exhibiting complete flatness. Here we present a large-scale analysis of binding geometry and physicochemical properties of all protein-protein interfaces available in the Protein Data Bank. Our analysis provides a comprehensive overview of the protein-protein interface landscape, including evidence that even for overall larger, more flat interfaces that utilise discontinuous interacting regions, small and potentially druggable pockets are utilised at binding sites.

Significance Statement

PPIs are promising but underutilised drug targets, partially due to the misconception that they are flat and featureless. This has been put into question with accumulation of structurally resolved complexes, thermodynamic characterisation and success of PPI modulators that have reached the market in recent years. The analysis presented here provides a detailed and more comprehensive understanding of the landscape of therapeutically relevant PPI interfaces, suggesting opportunities for more rational approaches to drug design. We show that while interfaces forming continuous segments make greater use of concavity, leading to being more tractable from a traditional druggability perspective, discontinuous interfaces are also amenable to modulation through allosteric or competitive inhibitors.

Introduction

Proteins are involved in most fundamental biological processes, including cell proliferation[1], signalling[2], host-pathogen interactions[3] and transport[4], via tightly coordinated and complex networks of interactions. Each protein will often interact through specific regions on their surface with several different protein partners. Given protein size and diversity, in humans, the proteome is estimated to be ~20,000, while the interactome over 650,000[5], with protein-protein interactions (PPIs) long been considered to offer a highly selective and tunable way to modulate protein activities and pathways.

Originally, interacting interface regions were considered to be large, hydrophobic, flat and featureless[6], leading to their characterisation as poor targets for the development of small molecule modulators. However, recent structural and thermodynamic characterisation[7] has allowed the classification of PPIs based on the nature of interacting partners, and further suggested that binding pockets at the interface may play important roles in molecular recognition and binding. However, due to the lack of understanding and complexity of PPI interface regions, this remains a challenging area.

While large compilations of PPI networks are important to elucidate which proteins interact with each other, they lack in-depth information of how those interactions occur. Despite a relatively small proportion of the interactome being covered by structural data, advances in experimental structure resolution and application of structural bioinformatics[8-10] add promising contributions to a more complete and broad structural characterisation of PPI interactions.

Here we report the results of a large-scale analysis for the structural landscapes of PPI interfaces based on 3D structures available in the Protein Data Bank (PDB)[11]. We investigate a range of geometric and physicochemical properties of over 55,000 PPI interfaces, including planarity, shape complementary, secondary structure content, solvent accessibility, use of concavity and identification of hotspots, across different classes of interfaces, and discussed implications for druggability.

Results

Protein-Protein Interface Properties: Analysis of the interface segmentation distribution of PPI interfaces within the PDB revealed that having up to 5 interface segments were the most prevalent, accounting for 70% of interfaces (Figure S1 and Table S1), with interactions involving peptides being predominantly single segmented. This allowed us to categorise the interfaces as either single (continuous) or multi-segmented (discontinuous). Figure S2 shows the distribution of planarity for interfaces of different segmentations and types. Single segmented interfaces were significantly more planar than multi segmented ones (Table S2) and, while there was no significant difference in segmentation between peptide-type interfaces, Non-identical pairs were significantly more planar than identical pairs with symmetric and non-symmetric interfaces. The former was the most planar among all interface types (Table S3).

Single and multi segmented interfaces were also largely composed of residues in loops and α-helices in their core and periphery regions (Figure S3, Figure S4, Table S4 and Table S5).

Loop residues dominated on average at smaller sides of single segmented interfaces, and on both sides of multi segmented interfaces, while β-sheet residues were significantly less prevalent in all interfaces. However, α-helix dominated in the interface cores of multi segmented interfaces and the larger sides of single segmented interfaces, but not in the smaller sides of single segmented interfaces. Loops were significantly more present in the interface peripheries of all segmentations of interfaces than α-helix, which in turn were significantly more present than β-sheet residues. With respect to secondary structure use by interface types, loops were more prevalent at identical non-symmetrical than α-helix, whereas there were no significant differences in α-helix and loop usage in identical symmetric interfaces. Peptides were significantly more looped than other interfaces, however, while the enzyme's interface regions of Enzyme-Peptide interfaces tended to be formed of loops, the protein interface regions of Protein-peptide interfaces were significantly more helical than unstructured. In the interface core, however, for peptides of Protein-peptide interfaces, α-helix made up a greater proportion of interface cores than α-helix overall, and helices were significantly more present in identical symmetric core residues than loops, as opposed to the opposite in the all interfaces. Loops were significantly more present in the interface peripheries of all interface types, followed by α-helix and β-sheets.

With respect to Normalised Interface Packing (NIP), single segmented interfaces were significantly more well-packed than multi segmented interfaces (Figure S5 and Table S6). Peptidic interfaces were most well packed, followed by identical pairs with non-symmetric interfaces and Non-identical pairs, which did not differ significantly in packing, and Identical pairs with symmetric interfaces (Table S7). Similar to NIP, Normalised Shape correlation (NSc) was significantly higher in single segmented interfaces than in multi segmented interfaces (Figure S6 and Table S8). Peptidic interfaces were the most complementary, however, Enzyme-peptide interfaces had significantly higher NSc values than Protein-peptide ones. Identical pairs with symmetric interfaces were the least complementary and Non-identical pairs and Identical pairs with non-symmetric interfaces were not significantly different from each other (Table S9).

The average buried surface area (BSA) was significantly higher for multi segmented interfaces than single segmented interfaces, by over 1,000 Å^2 (Figure S7). Single segmented interfaces used significantly greater proportions of interface core residues on their larger sides than either side of multi segmented interfaces (Tables S10 and Table S11). However, they utilised a significantly smaller proportion of interface core residues per interface on the smaller side of the interface than multi segmented interfaces, which differ significantly between smaller and larger side (Figure S8, Figure S9, Table S12 and Table S13).

Looking at the intermolecular interactions per 100 Å^2 BSA revealed interesting differences between the classes of interfaces. Figures S10-S11 and Tables S14-S45 show distributions of use of non-covalent contacts for PPI interfaces in the dataset, by interface segmentation and interface type, respectively. Single segmented interfaces were significantly enriched in VdW, hydrogen/polar, atom-ring interactions compared to interfaces with multiple segments, which showed to have significantly more ionic, hydrophobic, carbonyl, amide-ring and amide-amide interactions. With respect to types of interfaces, individual interaction types showed different variations. For some interface types, numbers of interactions per 100 Å^2 BSA matched those elucidated from analysing interactions by interface segmentation alone, such as peptidic interfaces making greater use of VDW clash, proximal, hydrogen/polar bonding,

weak hydrogen/polar bonding, hydrophobic, carbonyl, atom ring interactions. However, in other cases, variations between use of interactions were more interface type-dependent than segmentation-dependent. For example, there has been significantly more use of amide-amide interactions by Identical pairs with non-symmetric interfaces than any other interface type, while other interface types were not significantly different between interface types, with the exception of Protein-peptide interfaces, which made use of significantly fewer ionic interactions. Identical pairs with symmetric interfaces consistently made significantly lower or similar use of non-covalent interactions, with the exceptions of amide-amide, Carbon-PI and ionic interactions, compared to all other interface types.

Concavity Across Interfaces: Concave geometry of protein surfaces is implicated in the formation of surface regions suitable for the binding of small, potentially drug-like, molecules. The majority of observations indicated that both single and multi segmented interfaces made use of concavities over the whole interface surface, however, single segmented interfaces were bound significantly deeper on average, binding at a “groove” magnitude of concavity (Figure 1, Figure S12, Tables S46 and S47). By comparison, small-molecule natural product ligands occupy concavities of less than 5 Å with 60-95% of their atoms[12] (measured per-atom, rather than summarised by deepest value per-residue).

The importance of concavity on average and at the deepest level varied as the protein molecule size and interface size of the protomer increased (Figure S13) ($R=0.32$, p-value <0.05). Both single and multi segmented interfaces exhibited outliers with very large chain lengths. Single segmented interfaces also utilised significantly fewer interacting residues than multi segmented interfaces (Figure 2A and Tables S48), while each globular interface type was significantly different in number of interacting residues from one another (Table S49). No significant difference in the chain length for the two types of peptidic interfaces was observed, neither between identical pairs with symmetric and non-symmetric interfaces (Figure 2B, TableS S50 and S51). Notably, identical pairs with symmetric interfaces used significantly more residues than all the other types of interfaces.

Inspecting averaged concavity, showed that smaller protomers with smaller interfaces were more likely to utilise concavity on average (Figure S13, Tables S46 and S47). As protomer length increased, interfaces became overall flatter regardless of the number of interacting residues. With respect to deepest concavity utilised at interfaces, deep concavities (< 4 Å) were utilised by at least part of the interface for a majority of observations. However, interface deepest concavity tended to take less concave values for longer protomers with fewer interacting residues (Figure S13). Some exceptions to this trend were where longer protomers used deep concavities at their deepest, although the interacting region of these two large chains resembles more a peptidic interface.

Exploring use of Concavity: Looking more closely, we analysed how concavity at interfaces was used by individual residues. Residue utilisation of concavity, how well the residues of one side of each interface make use of the (sub-)pockets available to them on the partner protein, varied with the nearby formation of concavity on the binding partner protein (Figures S14 and S15). Here, single and multi segmented interfaces made use of concavity in both the core and periphery. For multi segment/globular interface categories, residues in the interface core were observed in bimodal distributions; a mode where the residue is bound deeply and using local concavity, and a mode where the residue is bound with varying degrees of local concavity on

the partner chain. Multi segment interfaces utilising discontinuous binding regions were not only larger than single segmented interfaces, but also less well packed and less complementary in shape compared to single segment interfaces. These observations suggest that single interacting segments make tight, selective interactions with their globular partner proteins, compared to looser interfaces in larger multi segmented complexes. Deepest average use of concavity was by peptide interface core residues, and peptide interface periphery residues occupied deeper concavities than Identical pairs with symmetric and non-symmetric interface core residues, which did not differ significantly.

The large proportion of interfaces that at their deepest occupied deep concavities (Figures S14 and S15) raised the hypothesis that both surfaces of PPI interfaces provide “anchoring” points for one another. Analysis of interfaces revealed that an “interlocking” phenomenon, where deep concavity utilised in the 0.5 Å to 2 Å range was complemented by reciprocal concavity use on the other side of the interface, existed in a greater proportion for multi segmented, globular interfaces, than for single segmented, peptidic interfaces (Figures 3 and S16). Helix residues bound significantly deeper than loop and sheet residues in single segmented interfaces given the same solvent accessibility, for multi segmented interfaces helices and sheets bound significantly deeper than loops, however, they were not significantly different from each other (Figure S17, Figure S18, Tables S52 and S53).

Concavity use by residues in different interfaces and solvent accessibility environments were also different by residue amino acid. Overall, bulkier residues used the most concavity than any residue environment in the interface core and periphery for single and multi segmented interfaces (Figure S19 and Table S54), and across the different interface types (Figure S20 and Table S55). Interestingly, minimum concavity values were consistently low across all interface types and solvent accessibilities, suggesting individual residue uses of concavity by each amino acid, in all residue environments.

Energetic Hot Spots: Hotspot density in different interface segmentations and types was calculated using mCSM-PPI to identify the number of hotspots per 100 $Å^2$ BSA (Figure S21). Single segment interfaces used significantly more hotspots per 100 $Å^2$ BSA than multi segmented interfaces. Interfaces involving peptides had the highest densities of hotspots and were significantly different between the two classes (Enzyme-peptide and Protein-peptide interactions) (Tables S56 and S57). For interactions involving globular proteins, Identical pairs with symmetric interactions used significantly more hotspots per 100 $Å^2$ BSA than the other two classes and identical pairs with non-symmetric interfaces utilised significantly fewer hotspots per 100 $Å^2$ BSA than any other interface type. Figures S22 and S23 illustrate the relationship between residue use of concavity, solvent accessibility, and energetic importance for each type of interface in the dataset. Overall, for residues originating from the more deeply bound sides of interfaces, there was no significant correlation between residue occupation of concavity and energetic importance (Pearson correlation coefficient R = -0.05). When separated by solvent accessibility, the correlations were R = 0.23 for interface core residues and R=0.02 for peripheral residues. Correlations of hotspots with use of concavity ranged from -0.04 to 0.25 for all interface types and environments (Figure S23).

Clustering of Orthosteric Sub-Pockets on PPI Interfaces: The anchor hypothesis of interaction proposes that initial, fast recognition between protomers is mediated by residues, usually from the smaller interacting partner, that bury a large portion (> 100 $Å^2$) of surface area

and adopt the same rotameric states when bound and unbound. We explored this concept using concavity as a metric for determining anchoring residues, in addition to solvent accessibility, which we define here as “enclosed” residues. Looking at the numbers of enclosed residues present in PPI interfaces (Figure S24) showed that around 80% of PPI interfaces had at least one enclosed residue. Enzyme-peptide interfaces exhibited the largest proportion of interfaces with at least one enclosed residue (93%), followed by Protein-Peptides (90%), Identical pairs with symmetric interface (88%), Identical pairs with non-symmetric interface (76 %), and non-Identical pairs (75%).

To explore how residues utilising concavity may be exploited for drug discovery, enclosed residues at PPI interfaces were clustered in 3D. These enclosed residue clusters represent pockets, or adjacent sub-pockets, that are demonstrably utilised by proteins at interfaces and thus have potential for orthosteric challenge with small-molecules. This revealed that 9,253 interfaces possessed enclosed residue clusters (16% of the dataset) (Figure 4). Protein-Peptide interfaces had the smallest proportion of interfaces with enclosed residue clusters (11%), followed by Identical pairs with non-symmetric interfaces (12%), non-identical pairs (12%), Enzyme-Peptides (13%), and identical pairs with symmetric interfaces with the highest proportion (26%).

The existence of small, buried protein-occupied pockets in larger, multi segment interfaces, consisting of clusters of multiple small-volume pockets may present opportunities for single residue sites to be competed for with fragments, which could be elaborated into interface competitive small molecules for transient interfaces where interface on/off kinetics could allow competitive inhibition. Geometric clustering of deeply bound and solvent inaccessible residues at interfaces revealed cases in the dataset that presented these dense clusters of enclosed residues, which were potentially occupying druggable pockets. However, the presence of such clusters is not an essential requisite for druggability, as evidenced by only one drugged PPI from the 2P2I dataset[13] possessing an enclosed cluster.

Discussion

In this study we explored the nature of PPI binding interfaces with respect to binding-mode geometry, interatomic interactions, and structural and energetic importance of interface residues. While often considered flat and featureless, we showed that while the majority of interfaces extracted from the PDB were indeed flat on average, many interfaces did utilise concavity at their deepest point, suggesting that an element of concavity is important for many protein-protein interactions. Peptidic interfaces and those utilising continuous binding regions at the interface made greater use of concavity on average, suggesting that these binding sites may be better defined with respect to potential exploitation in drug discovery. Depth may provide a way of improving encapsulation of a residue in smaller interfaces, as evidenced by the greater proportion of peptide interface core residues in Protein-Peptide interfaces using deeper binding modes, and making proportionally higher use of the local binding site space (complemented pockets) available in comparison to other residue environments. Our findings support the anchor hypothesis of many interfaces having deeply bound, solvent inaccessible and energetically important residues, which can be an important venue in drug discovery. We show that many interfaces provide concavity on both sides of the interface to support interactions.

We hypothesise that differences in interatomic contact usage by smaller, continuous interfaces compared to larger multi segmented interfaces may reflect differences in the nature of their recognition. As single segments tended to bind using more grooves than multi segmented PPIs, the significantly greater use of more specifically directional interactions, such as hydrogen bonding, by single segment interfaces may indicate an evolved imperative for the use of directional interactions to lock a segment into a deep binding site without requiring rearrangement of the globular binding partner. Conversely, for larger and multi segment interfaces, ionic interactions that may be involved in longer range electrostatic steering may contribute more to recognition where overall concavity is not present, and residues occupying concavities are less prevalent.

By analysing a large-scale dataset of structurally characterised PPIs from the PDB, we found that interfaces forming a continuous binding segment make greater overall use of protrusion into partner protein concavities on average than do globular discontinuous interactions. Deeply bound residues existed in a large proportion of all interactions and there was a relationship between depth and solvent accessibility depending on the continuity of the interface. Over 80% of interfaces utilised at least one deeply bound, solvent inaccessible residue, and over 16% of interfaces made use of multiple, small-volume sub-pockets of the kind bound by previously developed orthosteric PPI inhibitors.

We propose that while continuous binding sites that make use of concave binding modes overall may be more immediately tractable from a druggability perspective, there may be benefit in targeting globular protein interfaces with discrete, complemented sub-pockets, into which residue-sized small-molecule fragments could protrude. Through analysing the chemistry of interfaces as an aggregate property, summarising pairwise atomic interactions, we uncovered different chemical preferences between continuous and discontinuous binding sites, suggesting that single continuous segments require more specific directional interactions, whereas discontinuous interfaces burying larger surface areas rely more on aromatic sealing of the interface, and on electrostatic interactions. These discontinuous interfaces may be more amenable to target by allosteric or interface approaches. Our results move towards a better understanding of the features used at therapeutically relevant PPI interfaces, which can then be used on a more rational approach to drug design.

Materials and Methods

Data: Pairwise structures of interacting proteins were extracted from the PDB (accessed on 14 April 2021). Interactions with missing atoms at the interface, interfaces that overlapped with other interfaces (overlapping interfaces, where more than two protomers were bound together using the same residues, interfered with interpretation of concavity), interfaces where the product of the number of residues contributed by each protein partner was less than 25 and interfaces where less than 100 $Å^2$ was buried between the two proteins, were removed from the dataset. The latter two filters were used to remove interfaces where the chains did not make substantial contact[14]. To simplify large scale analysis, only the first model of NMR derived structures was considered.

A non-redundant set of PPI interfaces was generated by clustering interfaces first on whether the interacting pair of proteins was identical using CD-HIT at 95% identity cutoff[15] and

subsequently by clustering interactions involving identical protein chains based on the interface sequence. Here, we used the SequenceMatcher module, available in the *difflib* Python package, to compare short peptide sequences, with a similarity cutoff of 75%. Representative interface pairs for each cluster were chosen based on a structure quality score[14].

The final dataset of interfaces was partitioned by categorising interactions between globular proteins and protein-peptide interactions. The dataset consisted of 55,189 interfaces, of which 15,920 were Identical pairs with symmetric interface, 8,580 were Identical pairs with non-symmetric interface, 28,165 were Non-identical pairs, 1,702 were Protein-peptide interfaces, and 822 were Enzyme-peptide interfaces. Interactions between peptides and enzymes were separated from interactions with non-enzymatic proteins by identifying enzyme chains using the SIFTS cross-database mappings of the PDB to EC enzyme classification database[16], to differentiate enzyme-substrate and enzyme-inhibitor interactions that may involve active site cavities from non-catalytic site protein-peptide interfaces.

Interface Properties: Pairwise PPI interfaces consist of two interacting protein surfaces. Some properties of these interfaces, such as buried surface area, are property of the whole interface. However, other properties including binding depth belong to one side of the interactions. For the latter we conducted the analysis from the perspective of the smaller side of the interface (the side contributing the fewest residues; for example, the peptide in a protein-peptide interface), unless otherwise stated. Properties analysed included shape complementary, interface packing and planarity for whole interfaces. The shape correlation (Sc) measure uses interface region surface normal vectors to determine how well fit is the interface between two proteins[17]. However, in this work we used a more recent implementation which uses Delauney triangulation to calculate a Normalised Sc (NSc) and Interface Packing (NIP)[18]. Planarity of the interface was measure by using RMSD of interface residues Cα atoms from a least-squares fitted plane through the interface. The resulting planarity value, measured in angstroms (Å), is lower for more planar interfaces, and higher otherwise.

As for properties of protein residues, here we calculated type of secondary structure (α-helix, **β**-sheets and loop) using DSSP via Biopython[19], solvent accessibility was generated via NACCESS[20], non-covalent interactions were calculated using Arpeggio[21] and concavity was measured using the inaccessible probe radius ($R_{inaccess}$) value, in angstroms, calculated using Ghecom[12]. Concavity per residue was measured by using the deepest-bound atom's concavity value, while whole interface concavity was calculated via arithmetic mean of these deepest per-residue values across all interface residues.

Residues within 5 Å of any of the binding partner's protein atoms were considered to be part of the interface, and were further categorised as being core or periphery based on their solvent accessibility[14, 22]. Relative Solvent Accessibility (RSA) gives a measurement of burial from solvent that is comparable between residues of different volumes, and is used to determine which residues are buried in protein or interface cores. The categories used for residue solvent exposure are outlined in Table S58.

Energetically Important Interface Residues: The Ghecom measurement of concavity together with solvent accessibility was used to elucidate potential anchor residues from interface structure. Any residue that was solvent inaccessible with a residue minimum

concavity threshold of 4 Å or less were classified as enclosed residues. The DBSCAN density-based clustering algorithm[23] was used to geometrically cluster enclosed residues at interfaces to search for possible orthosteric pockets, defined by clusters of anchors.

In addition, $\Delta\Delta G^{Binding}$ values from mCSM-PPI[24] were used to perform computational alanine scanning of each interface, in order to determine the energetic importance of each binding residue. The threshold of $|\Delta\Delta G^{Binding}| > 1$ kcal/mol was then used to determine whether a residue was a hotspot or non-hotspot[25].

Statistical analysis: The one-way analysis of variance (ANOVA), as implemented in the stats module of SciPy[26], was used to compare distributions between different groups. Where ANOVA indicated significant differences between groups, we used Tukey's Honestly Significant Difference (Tukey's HSD) to categorise observations into their similar or different statistical significance using the Python module statsmodels[27].

Acknowledges

The authors wish to thank Dr Harry C. Jubb for his invaluable guidance and advice to this study. C.H.M.R is funded by a Melbourne Research Scholarship. This work was supported in part by the Medical Research Council (MR/M026302/1 to D.B.A. and D.E.V.P.); the National Health and Medical Research Council of Australia (GNT1174405 to D.B.A.), the Wellcome Trust (093167/Z/10/Z), and the Victorian Government's Operational Infrastructure Support Program.

References

1. Gao, J., W.X. Li, S.Q. Feng, Y.S. Yuan, D.F. Wan, W. Han, and Y. Yu, *A Protein-Protein Interaction Network of Transcription Factors Acting During Liver Cell Proliferation.* Genomics, 2008. **91**(4): p. 347-355.
2. Chuderland, D. and R. Seger, *Protein-Protein Interactions in the Regulation of the Extracellular Signal-Regulated Kinase.* Mol Biotechnol, 2005. **29**(1): p. 57-74.
3. Nicod, C., A. Banaei-Esfahani, and B.C. Collins, *Elucidation of Host-Pathogen Protein-Protein Interactions to Uncover Mechanisms of Host Cell Rewiring.* Curr Opin Microbiol, 2017. **39**: p. 7-15.
4. Paumi, C.M., J. Menendez, A. Arnoldo, K. Engels, K.R. Iyer, S. Thaminy, O. Georgiev, Y. Barral, S. Michaelis, and I. Stagljar, *Mapping Protein-Protein Interactions for the Yeast Abc Transporter Ycf1p by Integrated Split-Ubiquitin Membrane Yeast Two-Hybrid Analysis.* Mol Cell, 2007. **26**(1): p. 15-25.
5. Stumpf, M.P., T. Thorne, E. de Silva, R. Stewart, H.J. An, M. Lappe, and C. Wiuf, *Estimating the Size of the Human Interactome.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 6959-6964.
6. Jones, S. and J.M. Thornton, *Principles of Protein-Protein Interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13-20.
7. Jubb, H., T.L. Blundell, and D.B. Ascher, *Flexibility and Small Pockets at Protein-Protein Interfaces: New Insights into Druggability.* Prog Biophys Mol Biol, 2015. **119**(1): p. 2-9.

8. Stein, A., R. Mosca, and P. Aloy, *Three-Dimensional Modeling of Protein Interactions and Complexes Is Going 'Omics.* Curr Opin Struct Biol, 2011. **21**(2): p. 200-208.
9. Mosca, R., A. Ceol, and P. Aloy, *Interactome3d: Adding Structural Details to Protein Networks.* Nat Methods, 2013. **10**(1): p. 47-53.
10. Chakrabarti, P. and J. Janin, *Dissecting Protein-Protein Recognition Sites.* Proteins, 2002. **47**(3): p. 334-343.
11. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-242.
12. Kawabata, T., *Detection of Multiscale Pockets on Protein Surfaces Using Mathematical Morphology.* Proteins, 2010. **78**(5): p. 1195-1211.
13. Basse, M.J., S. Betzi, X. Morelli, and P. Roche, *2p2idb V2: Update of a Structural Database Dedicated to Orthosteric Modulation of Protein-Protein Interactions.* Database (Oxford), 2016. **2016**.
14. Bickerton, G.R., A.P. Higueruelo, and T.L. Blundell, *Comprehensive, Atomic-Level Characterization of Structurally Characterized Protein-Protein Interactions: The Piccolo Database.* BMC Bioinformatics, 2011. **12**: p. 313.
15. Li, W. and A. Godzik, *Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.
16. Velankar, S., P. McNeil, V. Mittard-Runte, A. Suarez, D. Barrell, R. Apweiler, and K. Henrick, *E-Msd: An Integrated Data Resource for Bioinformatics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D262-265.
17. Norel, R., S.L. Lin, H.J. Wolfson, and R. Nussinov, *Shape Complementarity at Protein-Protein Interfaces.* Biopolymers, 1994. **34**(7): p. 933-940.
18. Mitra, P. and D. Pal, *New Measures for Estimating Surface Complementarity and Packing at Protein-Protein Interfaces.* FEBS Lett, 2010. **584**(6): p. 1163-1168.
19. Cock, P.J., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M.J. de Hoon, *Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.
20. Hubbard, S.J., J.M.J.C.P. Thornton, Department of Biochemistry, and U.C.L. Molecular Biology, *Naccess.* 1993. **2**(1).
21. Jubb, H.C., A.P. Higueruelo, B. Ochoa-Montano, W.R. Pitt, D.B. Ascher, and T.L. Blundell, *Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures.* J Mol Biol, 2017. **429**(3): p. 365-371.
22. Hubbard, T.J. and T.L. Blundell, *Comparison of Solvent-Inaccessible Cores of Homologous Proteins: Definitions Useful for Protein Modelling.* Protein Eng, 1987. **1**(3): p. 159-171.
23. Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. in *kdd*. 1996.
24. Pires, D.E., D.B. Ascher, and T.L. Blundell, *Mcsm: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures.* Bioinformatics, 2014. **30**(3): p. 335-342.
25. Ascher, D.B., H.C. Jubb, D.E. Pires, T. Ochi, A. Higueruelo, and T.L. Blundell, *Protein-Protein Interactions: Structures and Druggability*, in *Multifaceted Roles of Crystallography in Modern Drug Discovery*. 2015, Springer. p. 141-163.

26. Virtanen, P., R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, I. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, and C. SciPy, *Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python.* Nat Methods, 2020. **17**(3): p. 261-272.
27. Seabold, S. and J. Perktold. *Statsmodels: Econometric and Statistical Modeling with Python*. in *Proceedings of the 9th Python in Science Conference*. 2010. Austin, TX.

Figures

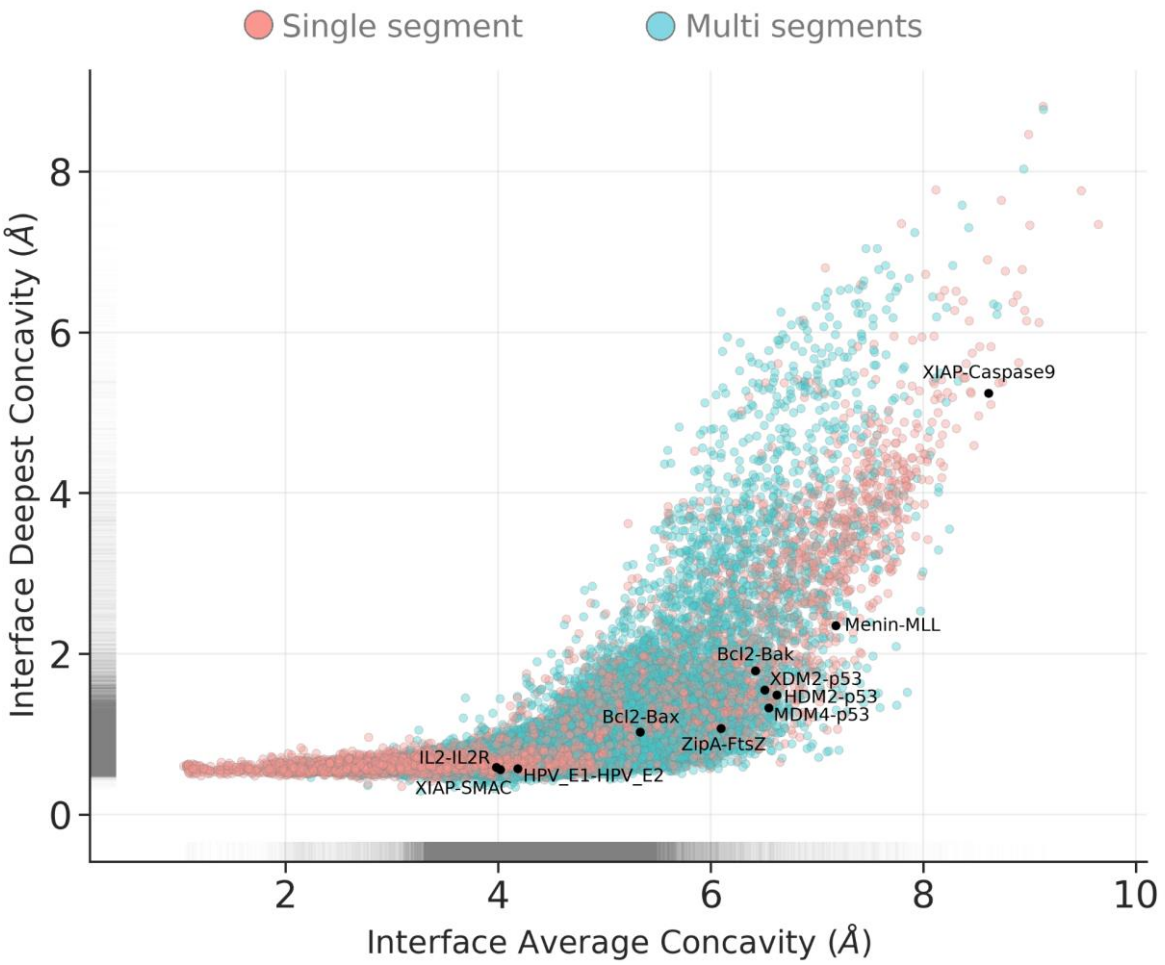


Figure 1 - Point and 2D density distributions of occupation of concavity at PPI interfaces, on average and at deepest point. Each point represents the smaller side of one interface from the non-redundant set of non-overlapping PPI interfaces. Concavity is as measured by Ghecom, representing the smallest spherical probe size that was able to enter a space around the partner protein's surface (where smaller values represent deeper binding). Interfaces are coloured by segmentation, and PPI interfaces from 2P2I dataset for which small-molecule inhibitors have been developed are overlaid as black points and are labelled.

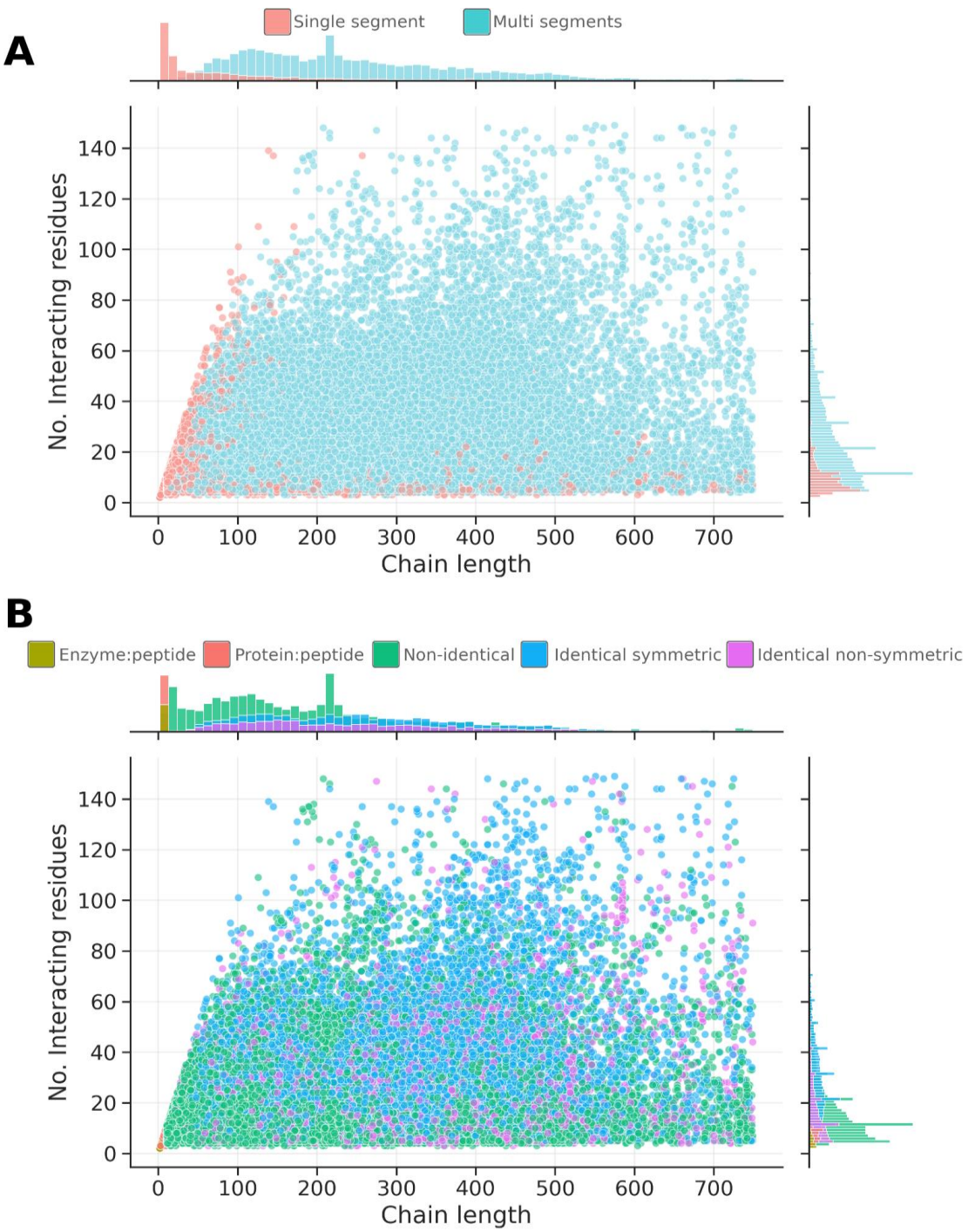


Figure 2 - 2D density distributions showing interface classifications by chain length and size of interacting surfaces. Density distributions are shown at a single density level for interfaces by A) segmentation and B) interface type.

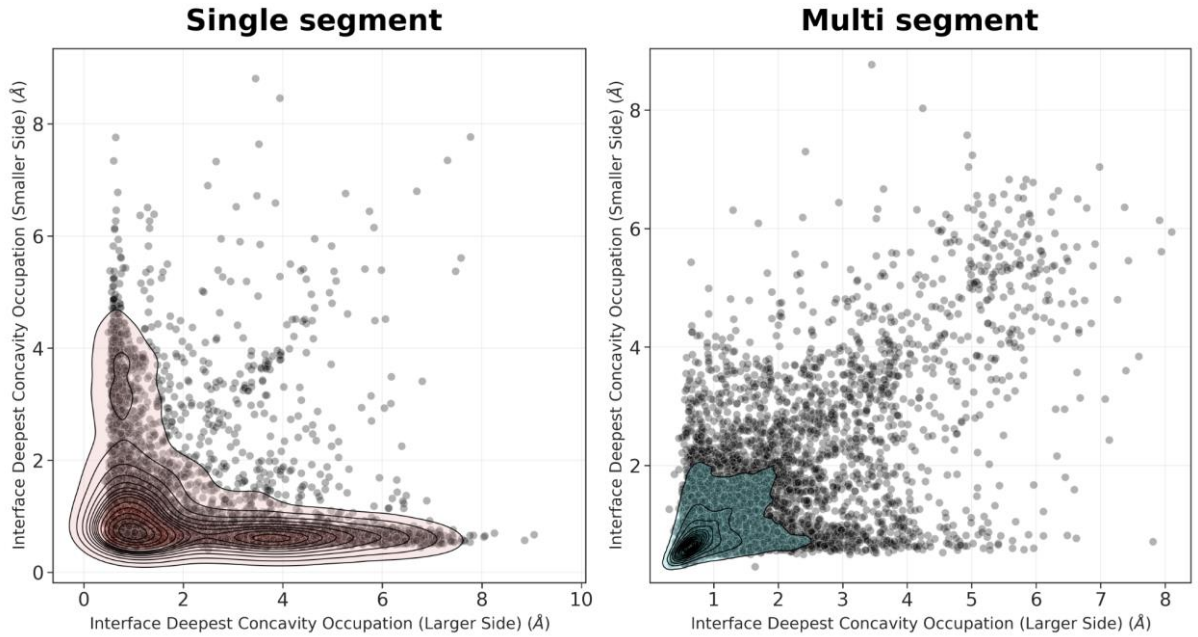


Figure 3 - Point and 2D density distributions of deepest concavity occupation on the larger and smaller sides of PPI interfaces. Concavity is as measured by Ghecom, representing the smallest spherical probe size that was able to enter a space sound the partner protein's surface (where smaller values represent deeper binding). Density distributions are coloured by interface segmentation.

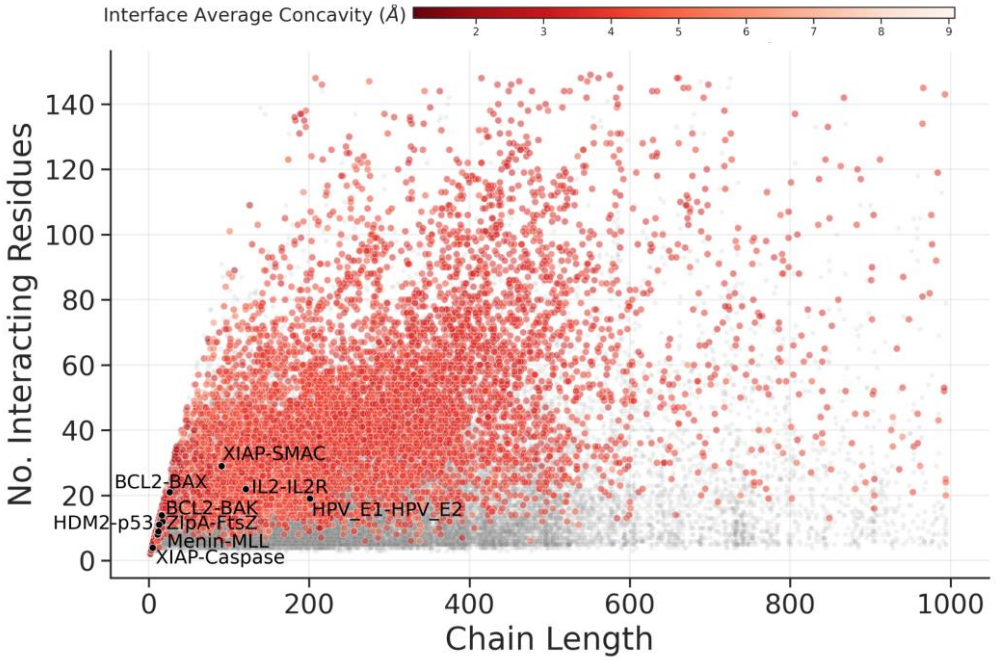


Figure 4 - Elucidating potential orthosteric binding pockets utilised by PPI protein partners, by clustering deeply bound, solvent inaccessible interface residues. The distribution of protein partner chain length as compared to binding site size is shown as gray points overlaid with coloured circles representing interfaces for which clusters of enclosed residues were found. Interfaces from the 2P2I set for which small-molecule inhibitors have been designed are overlaid as circles and labelled. 2P2I interfaces for which an enclosed residue cluster was found are marked by coloured circles.

PNAS
www.pnas.org

Supplementary Information for

Structural Landscapes of PPI Interfaces

Carlos H. M. Rodrigues [1,2,3], Douglas E. V. Pires [1,2,4], Tom L. Blundell [5], David B. Ascher [1,2,3,5,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria
[2] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria
[3] School of Chemistry and Molecular Biosciences, Bio21 Institute, University of Queensland, Brisbane, Victoria
[4] School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria
[5] Department of Biochemistry, University of Cambridge, Cambridge, UK

[*]To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au.

This PDF file includes:

Figures

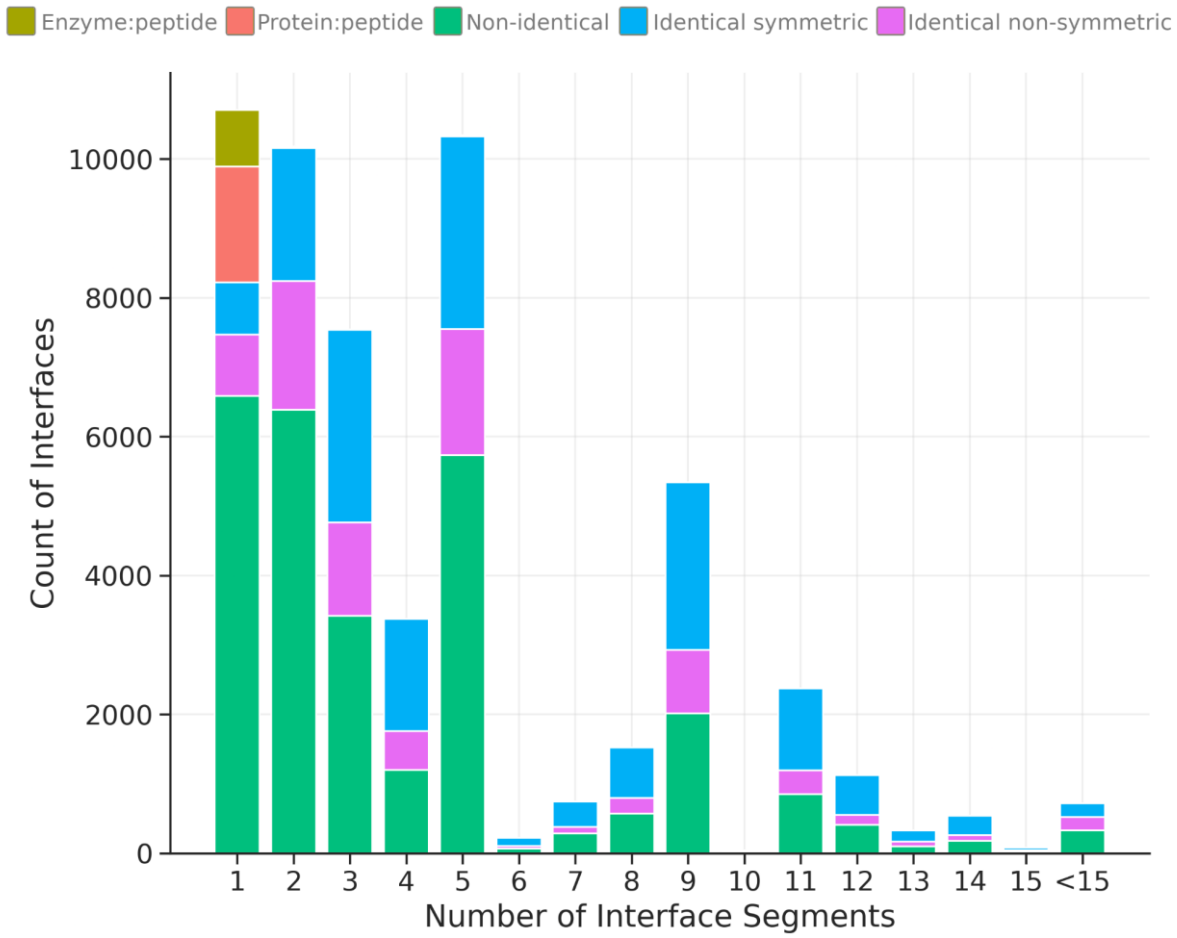


Figure S1 - Histogram distribution of interface binding continuity for PPI interfaces. Binding continuity is shown for the smaller side (fewest interacting residues) of each pairwise PPI interface in our non-redundant set. Interfaces are distinguished by interaction type.

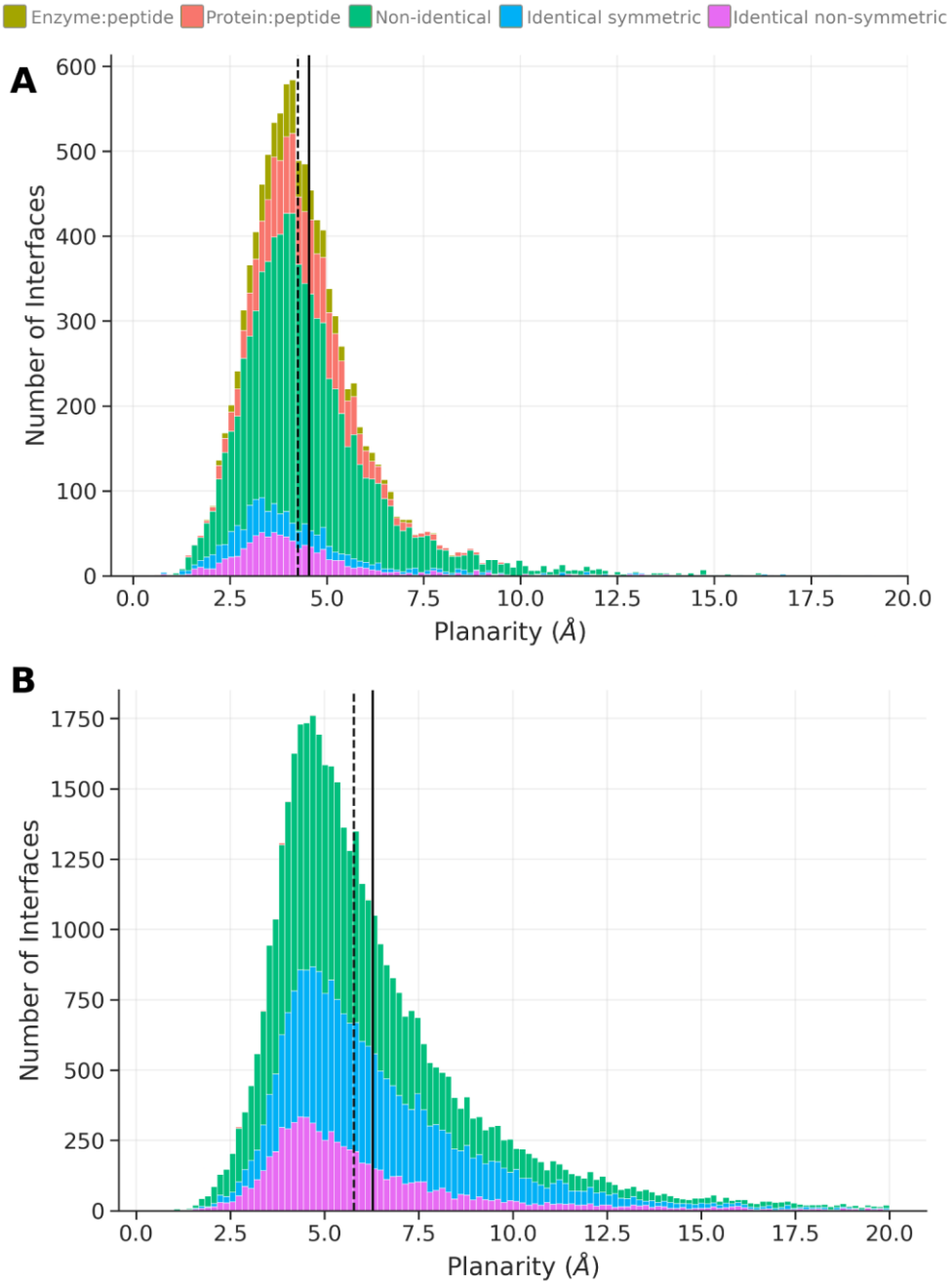


Figure S2 - Histogram distributions of interface planarity for single and multi segmented interfaces. Histograms are stacked and filled based on the type of interface. A) displays the distribution of interface planarity for single segmented interfaces and B) for discontinuous interfaces. Solid line indicates distribution mean and dashed line shows the geometrical mean.

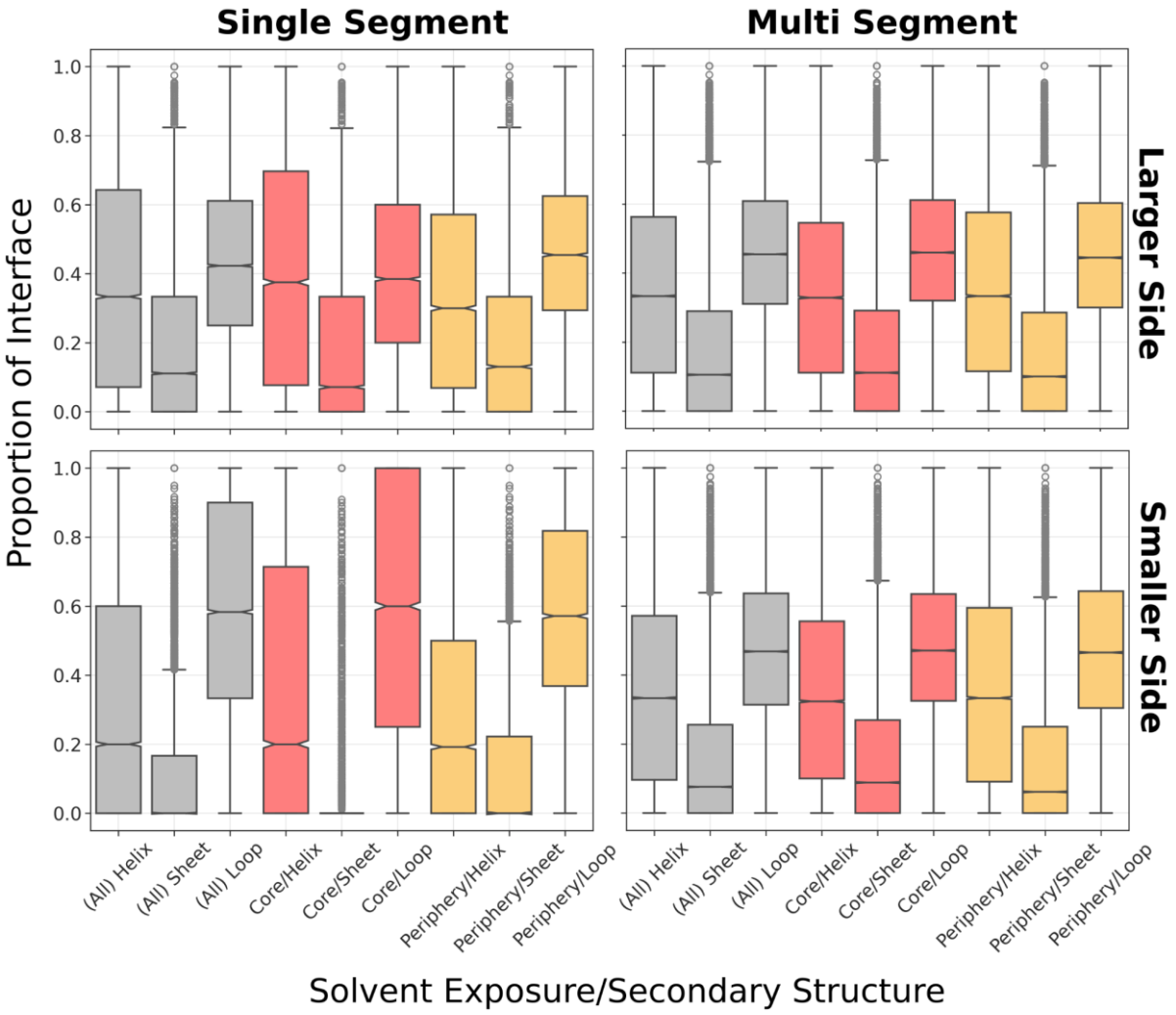


Figure S3 - Boxplot distributions of interface proportions of secondary structure types, by interface segmentation. Gray boxes indicate proportions of interfaces with a given secondary structure type. Coloured boxes represent proportions of interfaces with a given secondary structure by interface solvent exposure type.

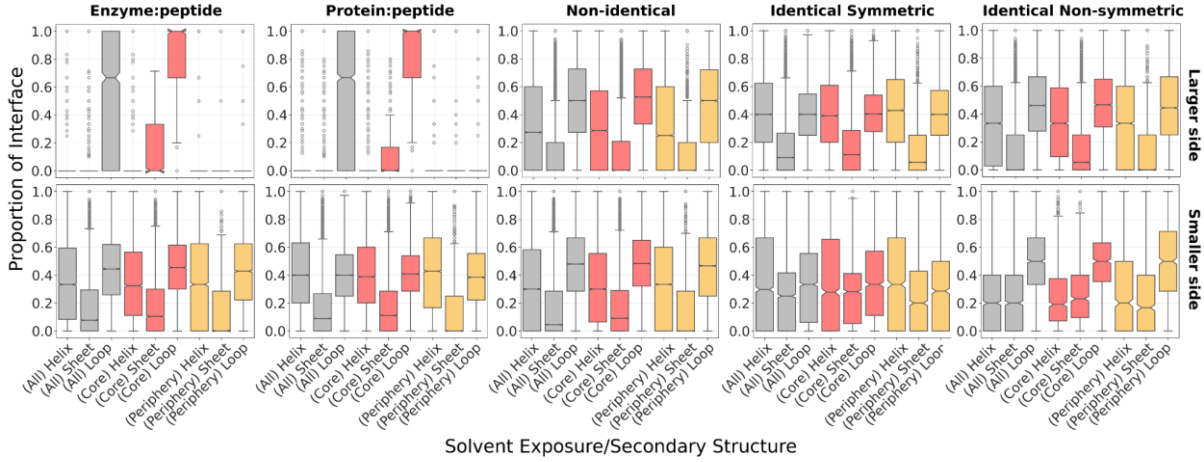


Figure S4 - Boxplot distributions of interface proportions of secondary structure types, by interface segmentation and interface type. Gray boxes indicate proportions of interfaces with a given secondary structure type. Coloured boxes represent proportions of interfaces with a given secondary structure by interface solvent exposure type.

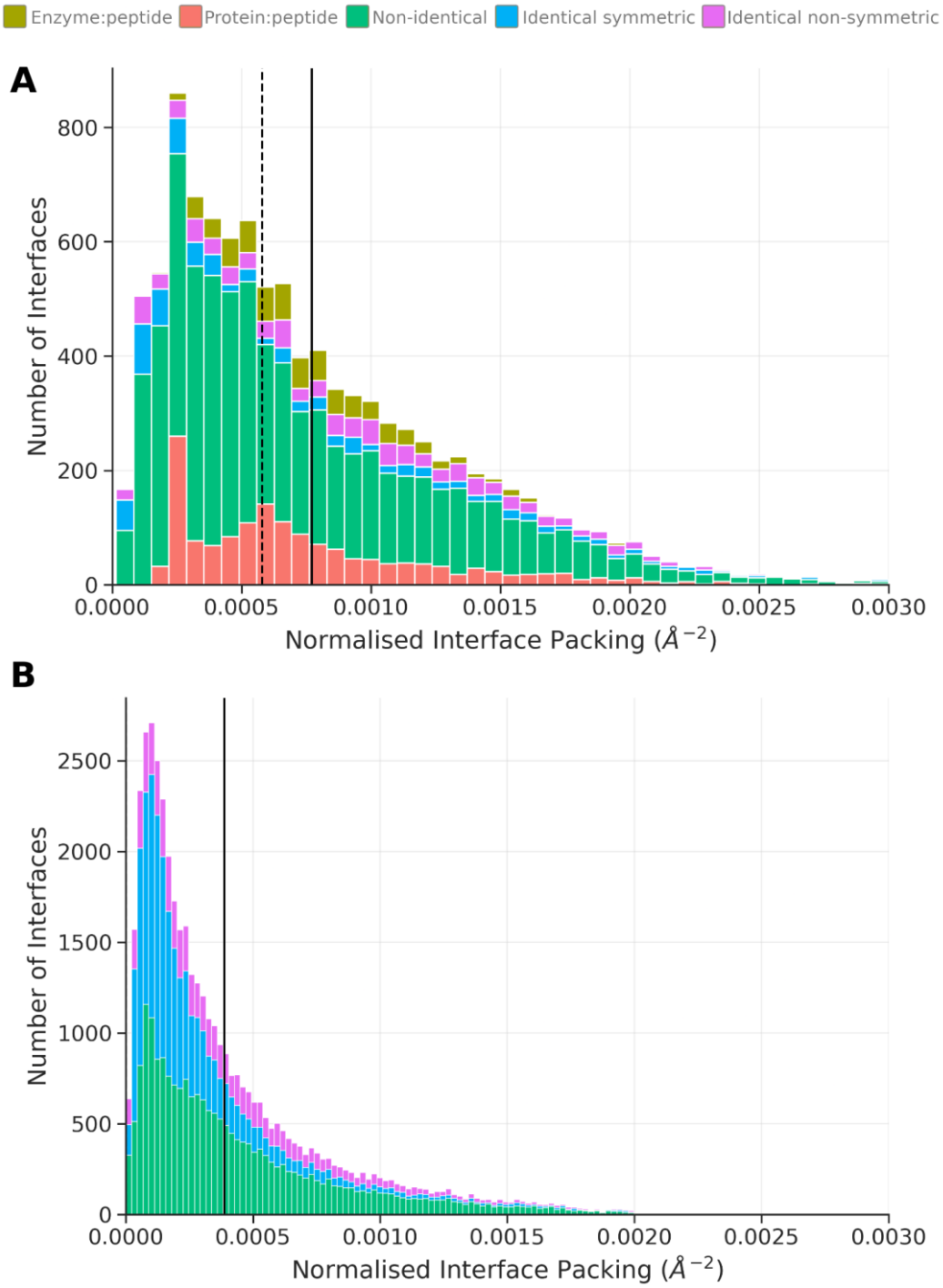


Figure S5 - Histogram distributions of interface normalised interface packing (NIP) for single and multi segmented interfaces. Histograms are stacked and filled based on the type of interface. A) displays the distribution of interface packing for single segmented interfaces and B) for discontinuous interfaces. Solid line indicates distribution mean and dashed line shows the geometrical mean.

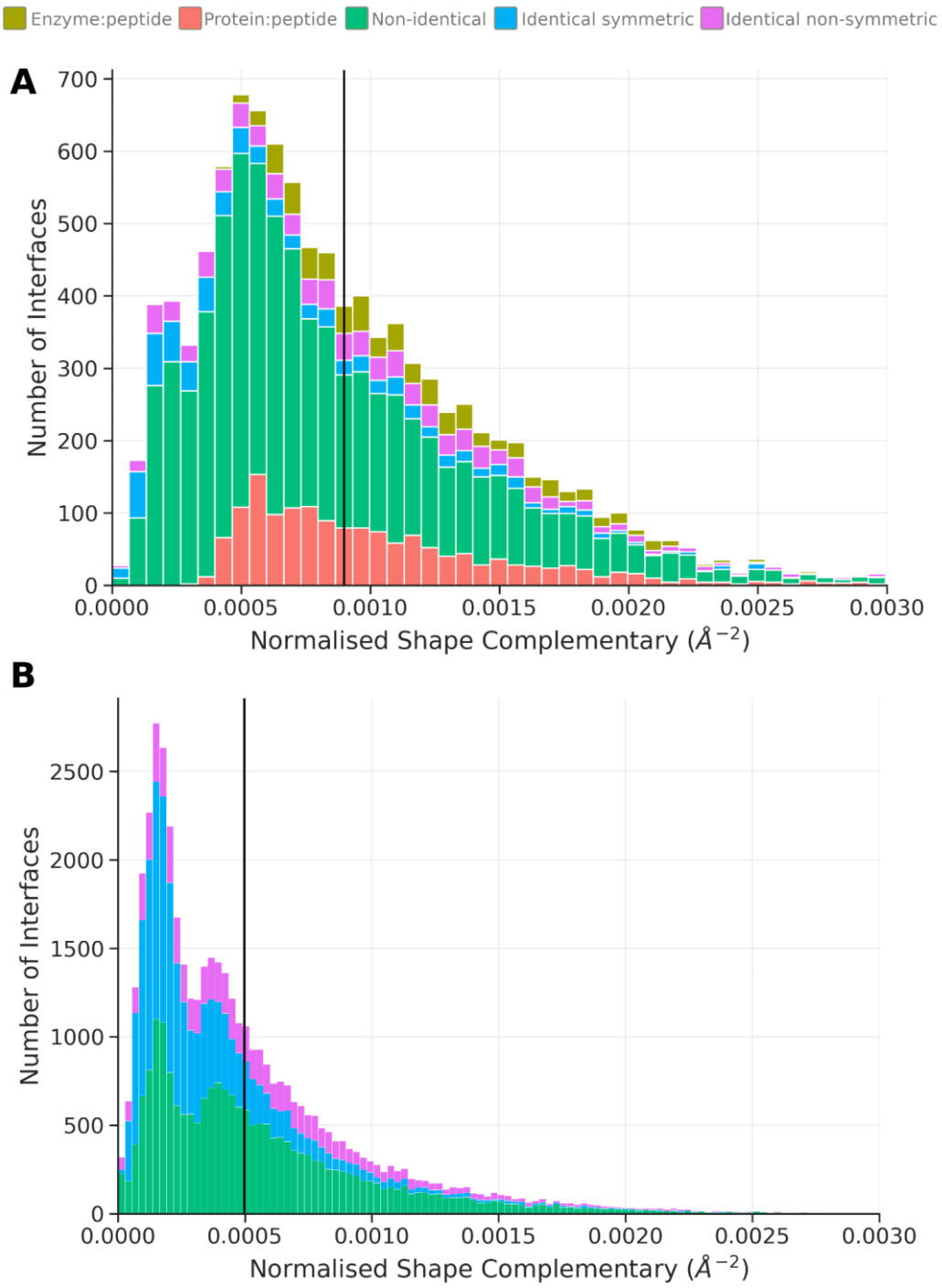


Figure S6 - Histogram distributions of interface normalised shape correlation (NSc) for single and multi segmented interfaces. Histograms are stacked and filled based on the type of interface. A) displays the distribution of interface shape complementarity for single segmented interfaces and B) for discontinuous interfaces. Solid line indicates distribution mean and dashed line shows the geometrical mean.

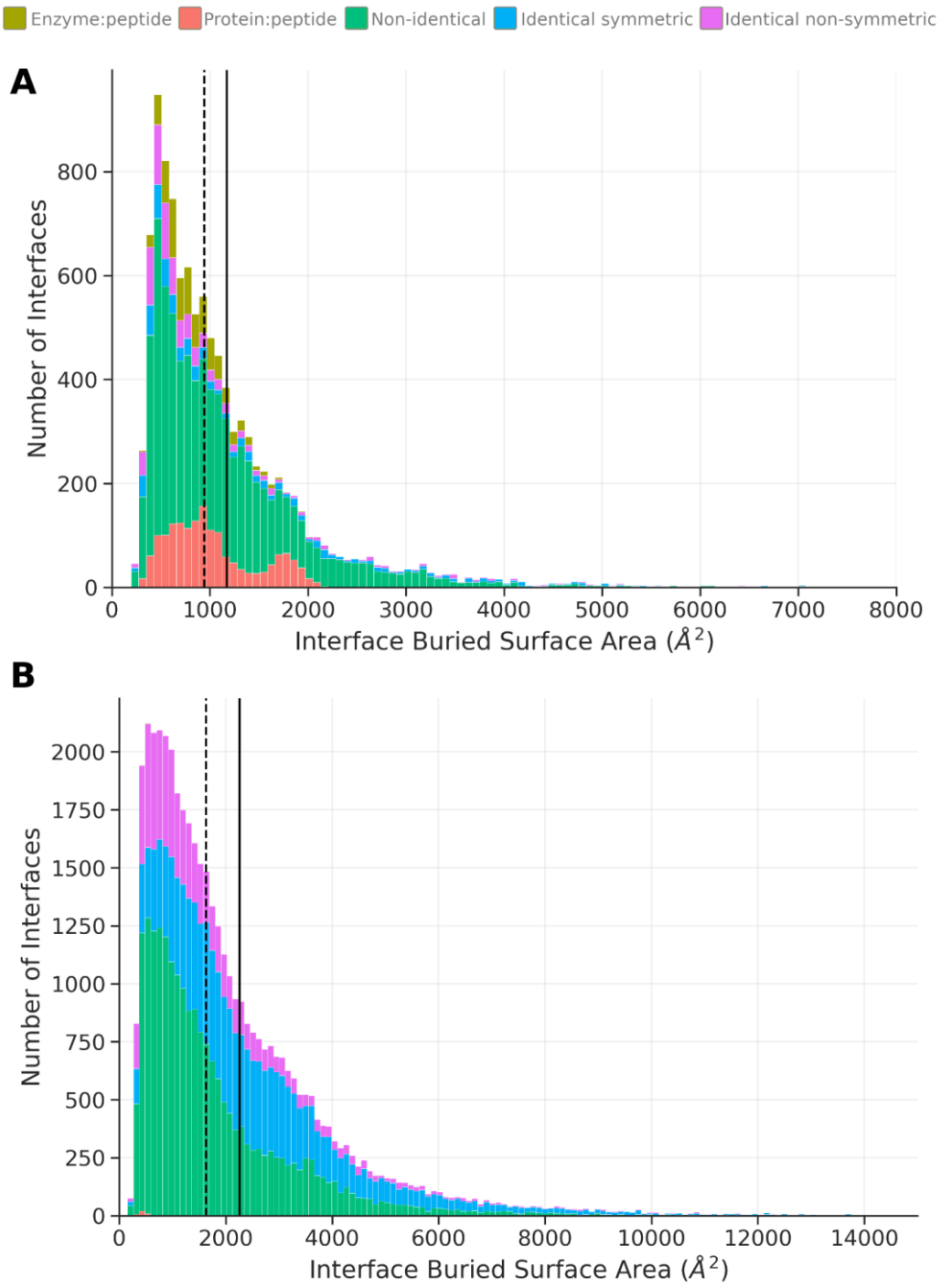


Figure S7 - Histogram of buried surface area distributions for single and multi segmented interfaces. Histograms are stacked and filled by type of interface. A) displays the distribution of interface shape complementarity for single segmented interfaces and B) for discontinuous interfaces. Solid line indicates distribution mean and dashed line shows the geometrical mean.

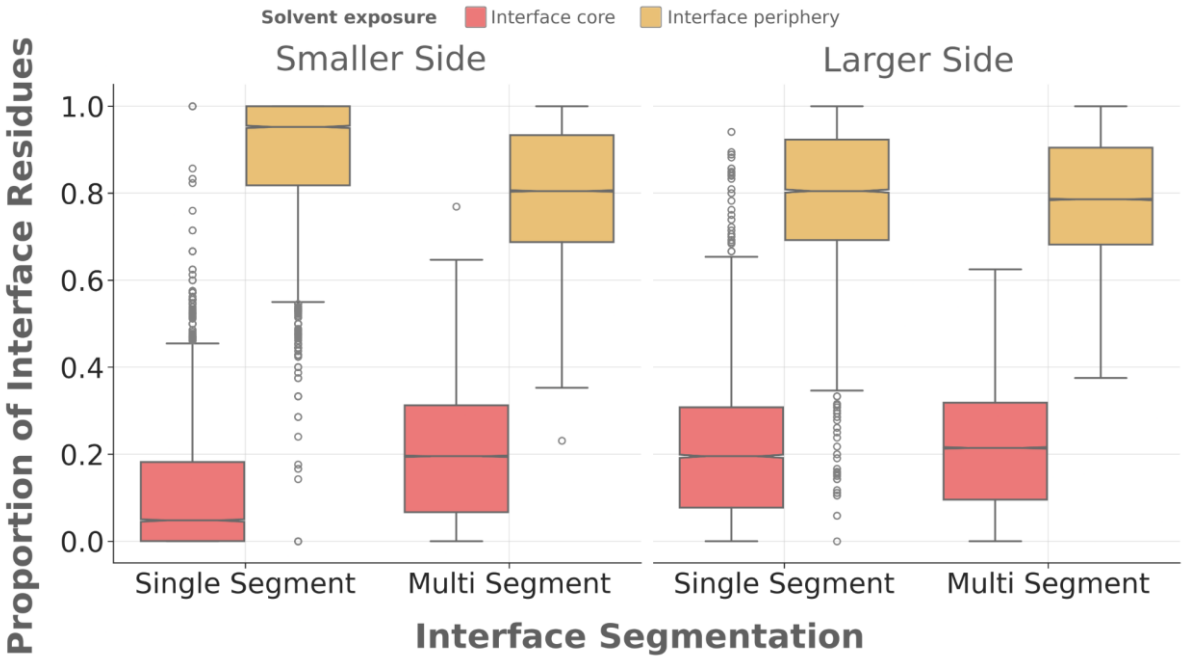


Figure S8 - Boxplot distributions of interface proportions of residues that were core or periphery, by interface segmentation. Outliers are shown as translucent gray circles. “Smaller side” refers to the side of the pairwise interface with fewer interacting residues than the “Larger side”.

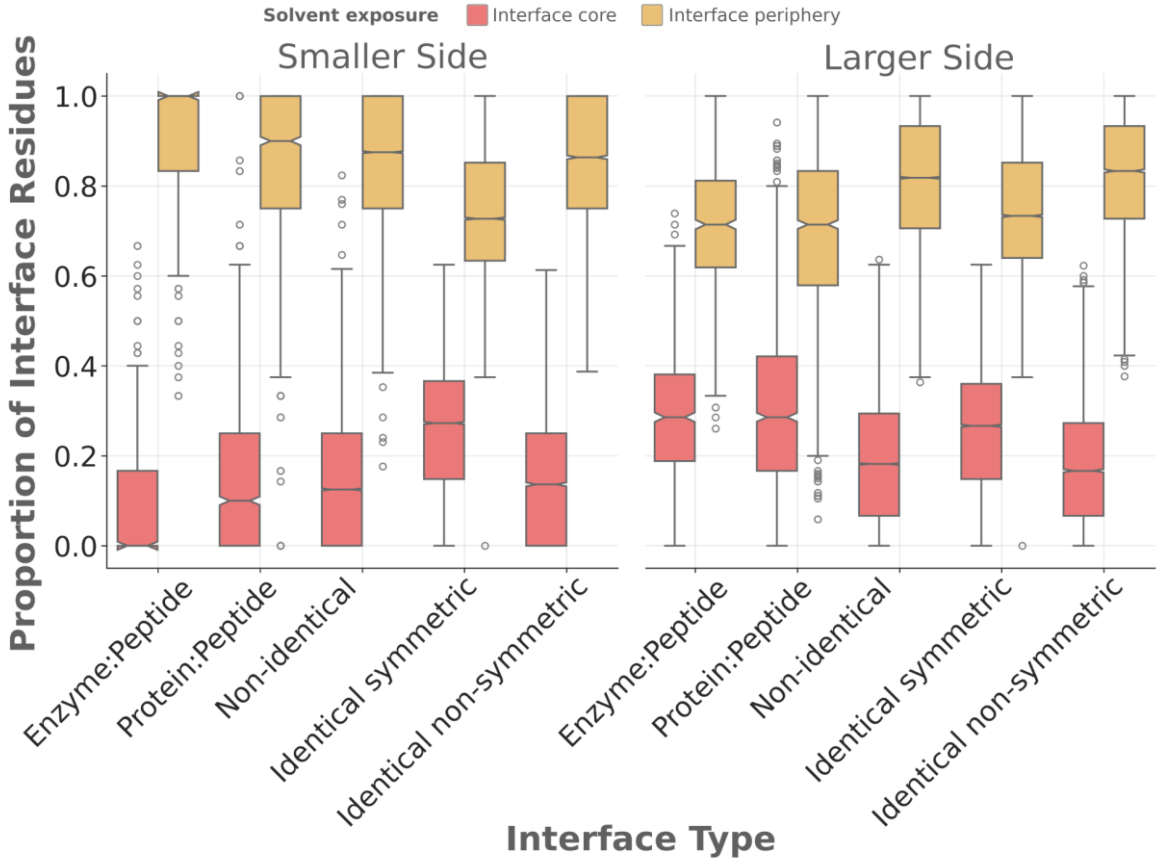


Figure S9 - Boxplot distributions of interface proportions of residues that were core or periphery, by interface segmentation and interface type. Outliers are shown as translucent gray circles. “Smaller side” refers to the side of the pairwise interface with fewer interacting residues than the “Larger side”.

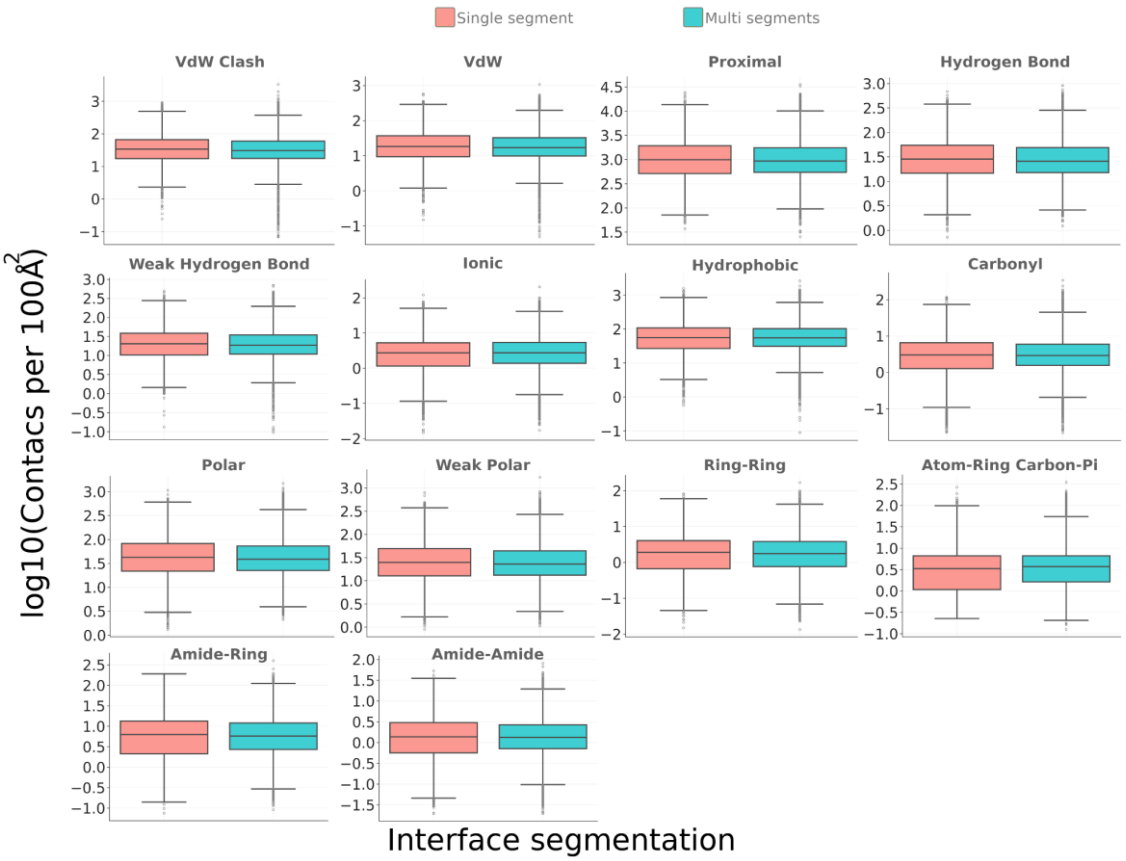


Figure S10 - Boxplot distributions of Arpeggio structural interactomics analysis of non-covalent interactions per 100Å² of PPI interfaces, comparing interfaces by segmentation. Outliers are shown as gray circles.

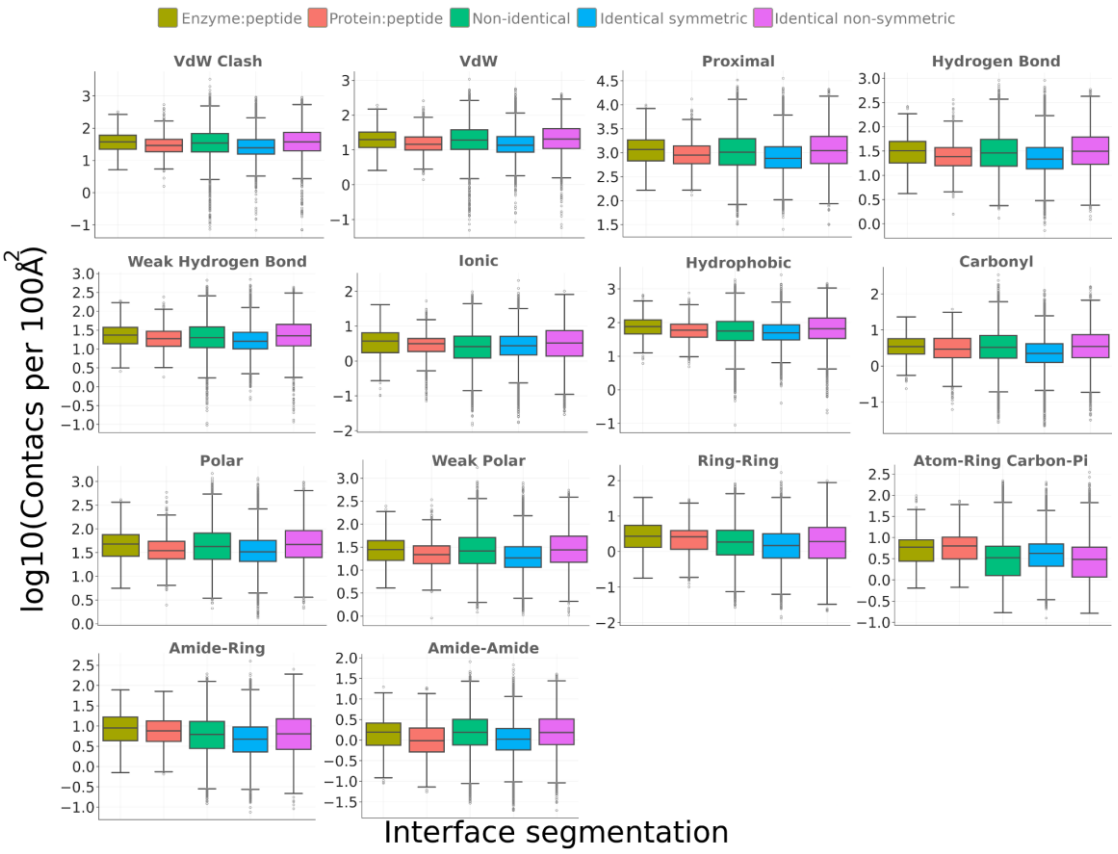


Figure S11 - Boxplot distributions of Arpeggio structural interactomics analysis of non-covalent interactions per 100Å² of PPI interfaces, comparing interfaces by interface type. Outliers are shown as gray circles.

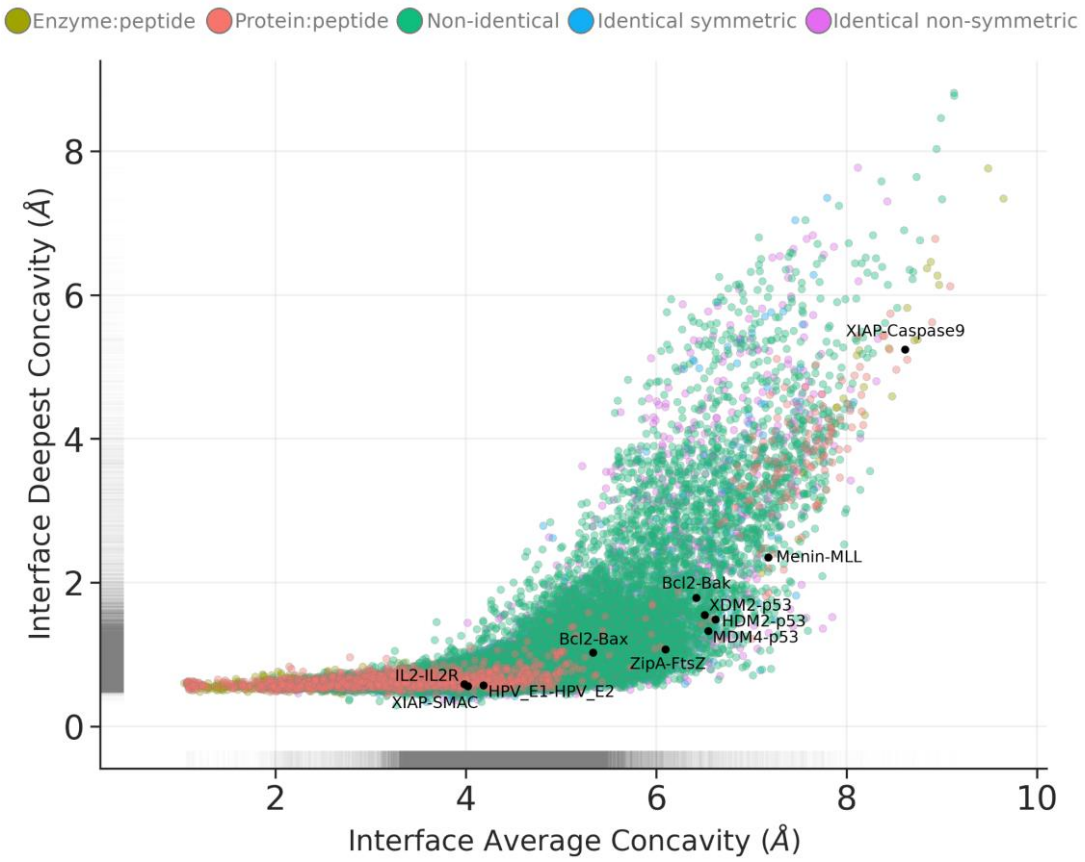


Figure S12 - Point and 2D density distributions of occupation of concavity at PPI interfaces, on average and at deepest. Each point represents the smaller side of one interface from the non-redundant set of non-overlapping PPI interfaces. Concavity is measured by Ghecom, representing the smallest spherical probe size that was able to enter a space around the partner protein's surface 9where smaller values represent deeper binding). Interfaces are coloured by interface type, and PPI interfaces from the 2P2I dataset for which small-molecule inhibitors have been developed are overlaid as black points and are labelled.

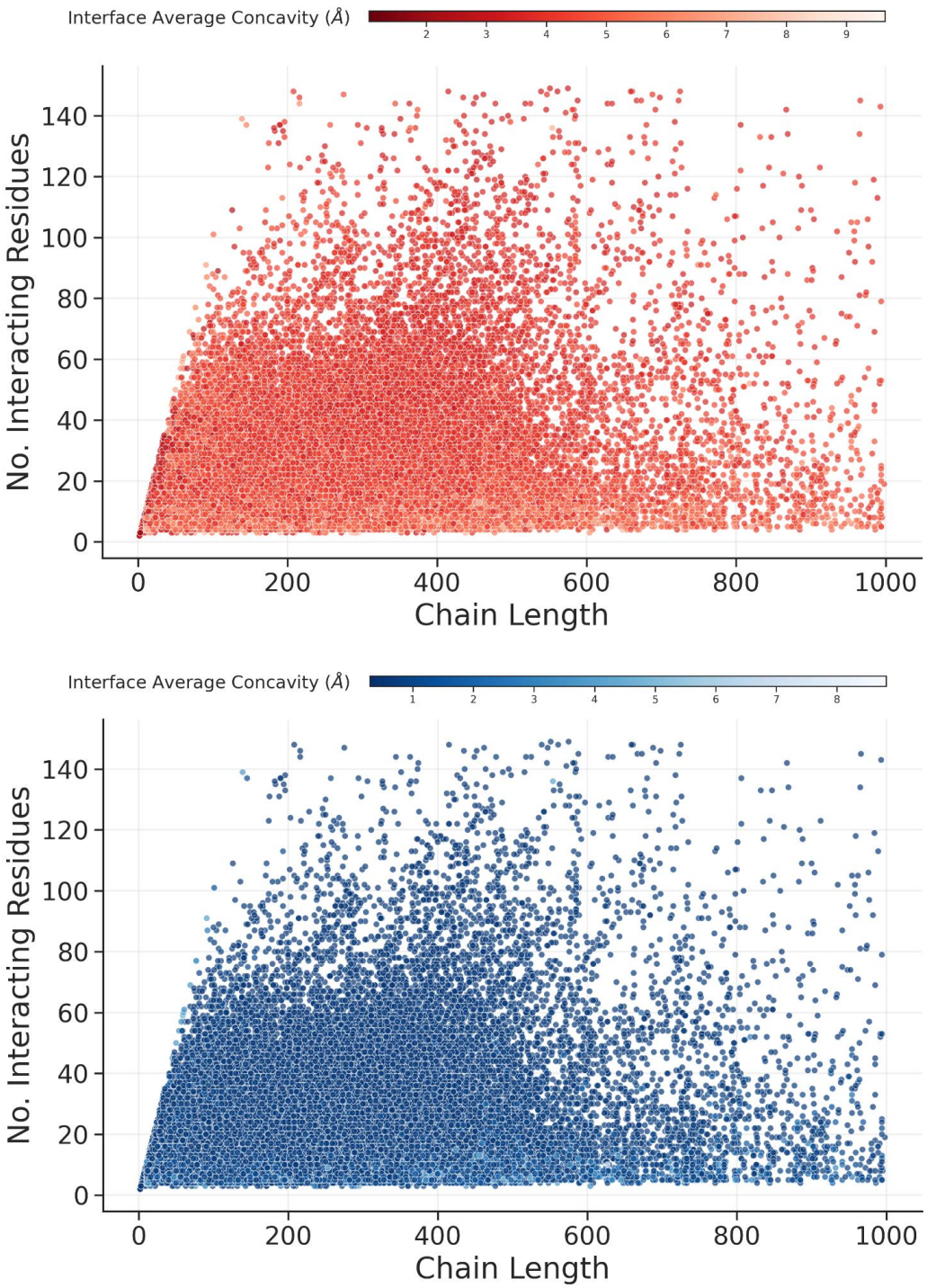


Figure S13 - Summary plots showing exploitation of concavity by PPI interfaces as the size of interacting protomers and their interacting surface varied. Each coloured block represents the arithmetic mean of A the average (red) or B the deepest (blue) concavity exploited by interfaces within the plot space covered by the block.

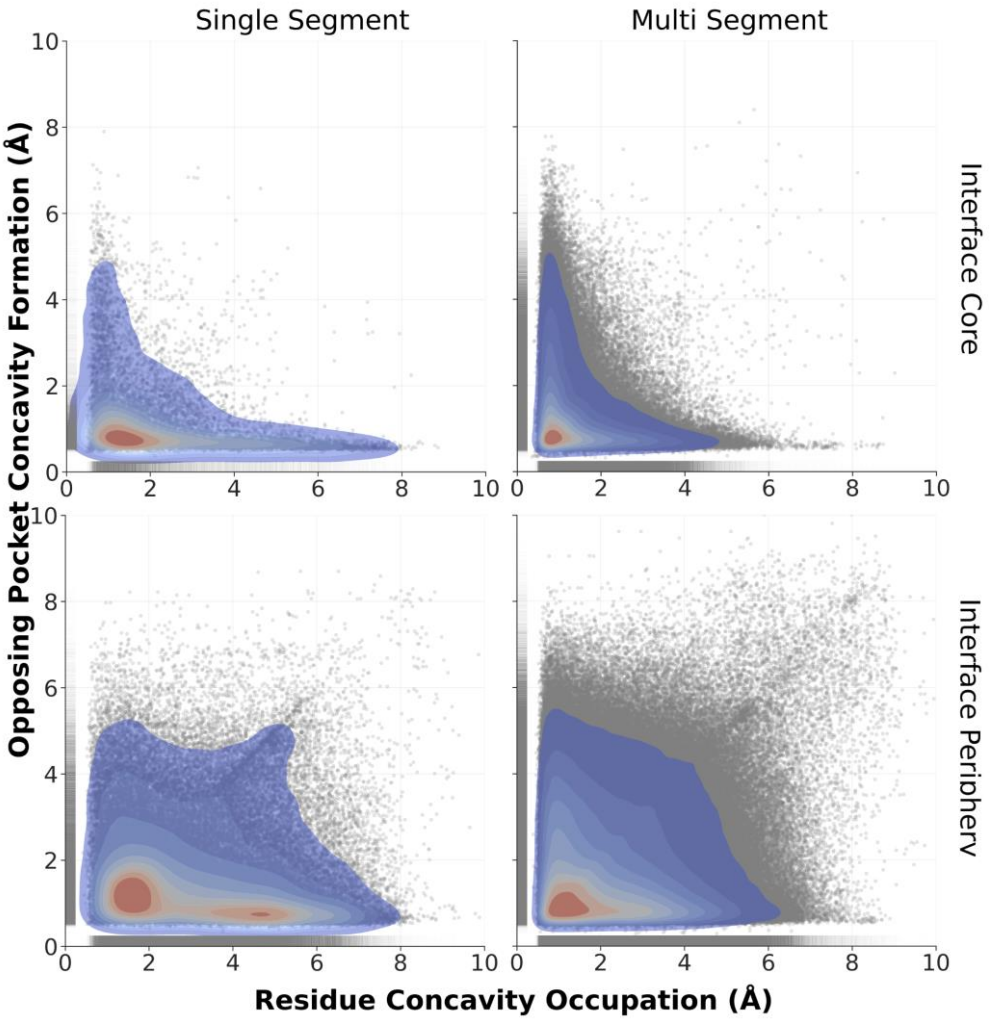


Figure S14 - Distribution of binding mode and extent of local concavity of residue in different interface segmentations. The use of concavity interface residues from the shorter chain of interactions of each class is shown on the abscissa. The depth of concavity formation by the deepest surface atom of the partner protein withing 5Å of the abscissa residue is shown on the ordinate. Each open circle represents a residue, and the 2D density distribution is shown where red, orange, yellow and blue represent areas of higher through lower point density, respectively.

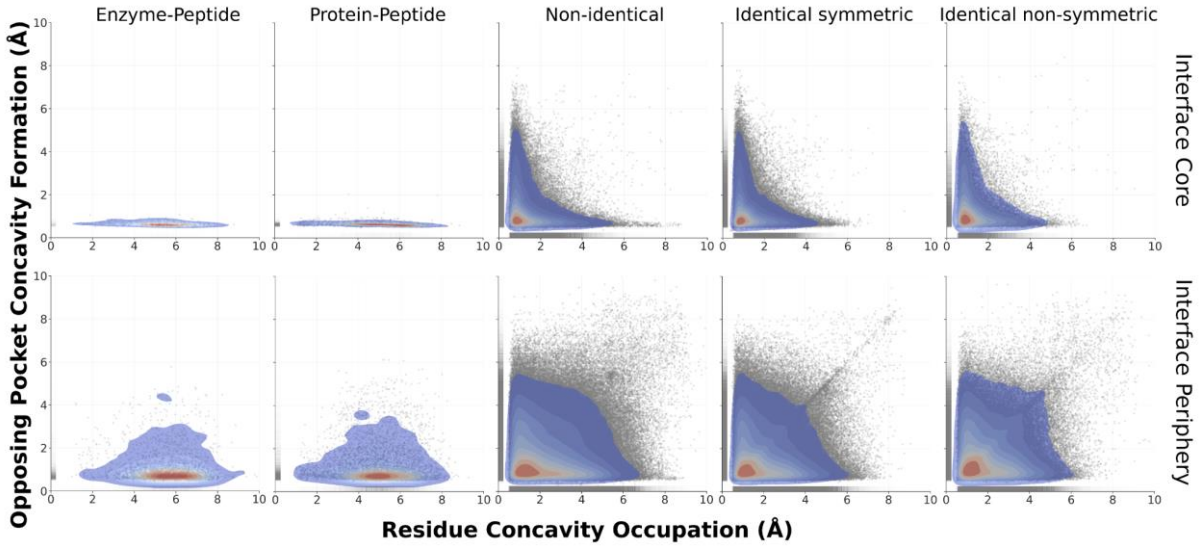


Figure S15 - Distribution of binding mode and extent of local concavity for residues in different interface classes. The use of concavity interface residues from the shorter chain of interactions of each class is shown on the abscissa. The extent of concavity formation by the deepest surface atom of the partner protein within 5Å of the abscissa residue is shown on the ordinate. Each open circle represents a residue, and the 2D density distribution is shown where red, orange, yellow and blue represent areas of higher through lower point density, respectively.

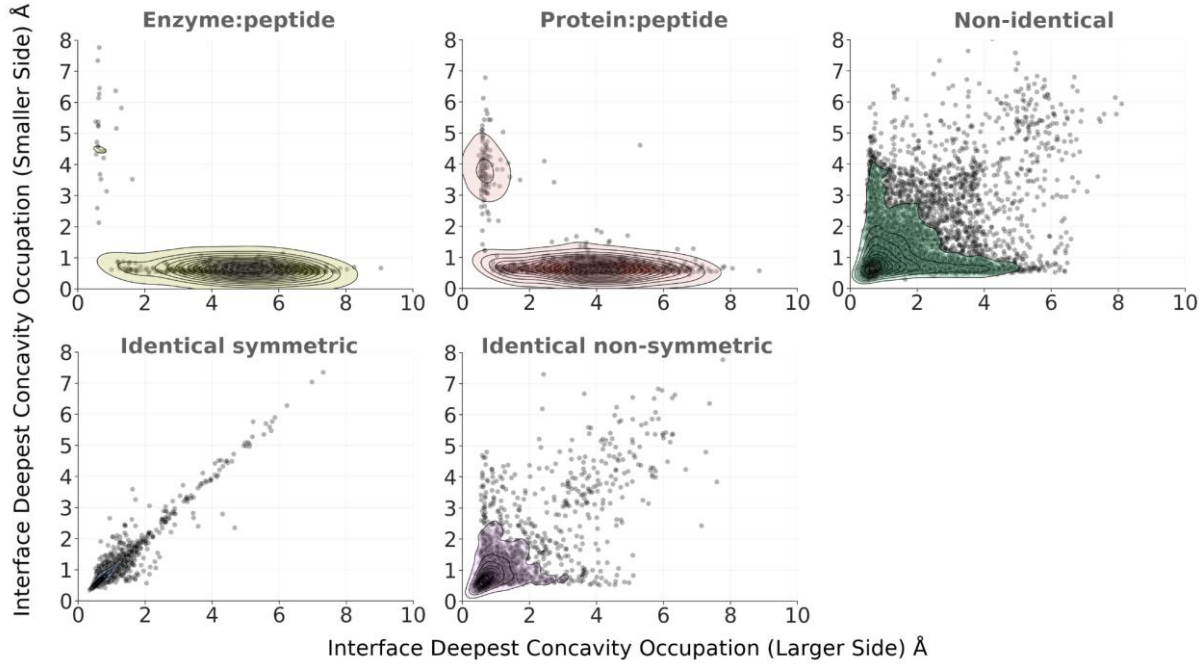


Figure S16 - Point and 2D density distributions of deepest concavity occupation on the larger and smaller sides of PPI interfaces. Concavity is measured by Ghecom, representing the smallest spherical probe size that was able to enter a space around the partner protein's surface (where smaller values represent deeper binding). Density distributions are coloured according to interface type.

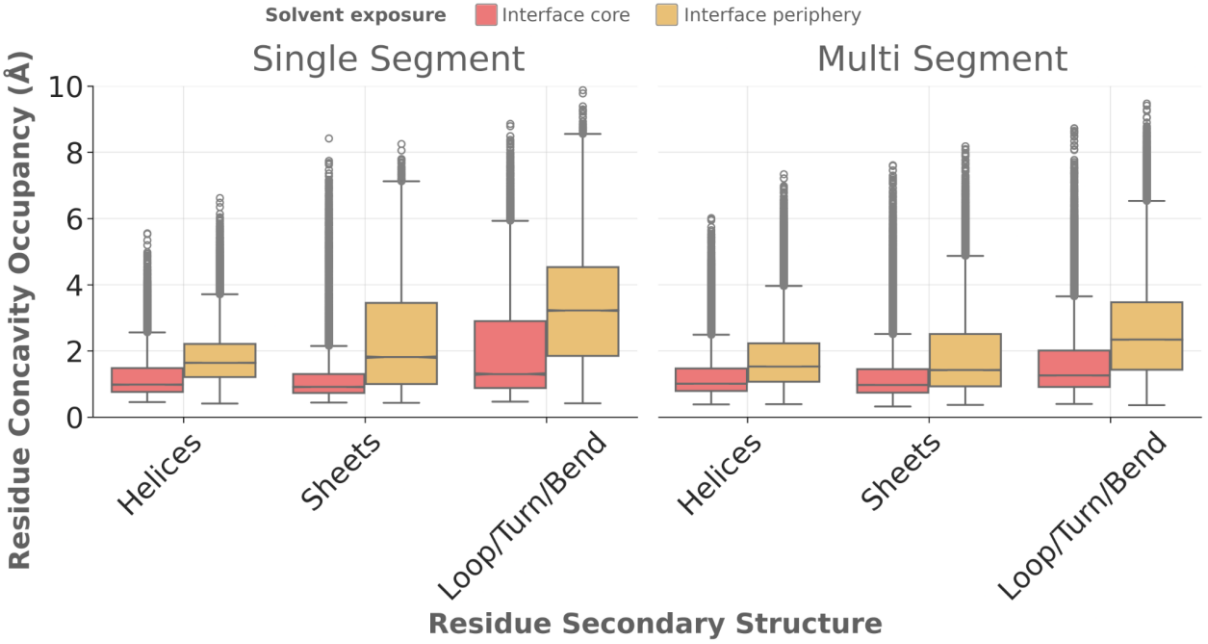


Figure S17 - Boxplot distributions of interface residue use of concavity by secondary structure, solvent exposure and interface segmentation. Concavity is measured by Ghecom, representing the

smallest spherical probe size that was able to enter a space around the partner protein's surface (where smaller values represent deeper binding).

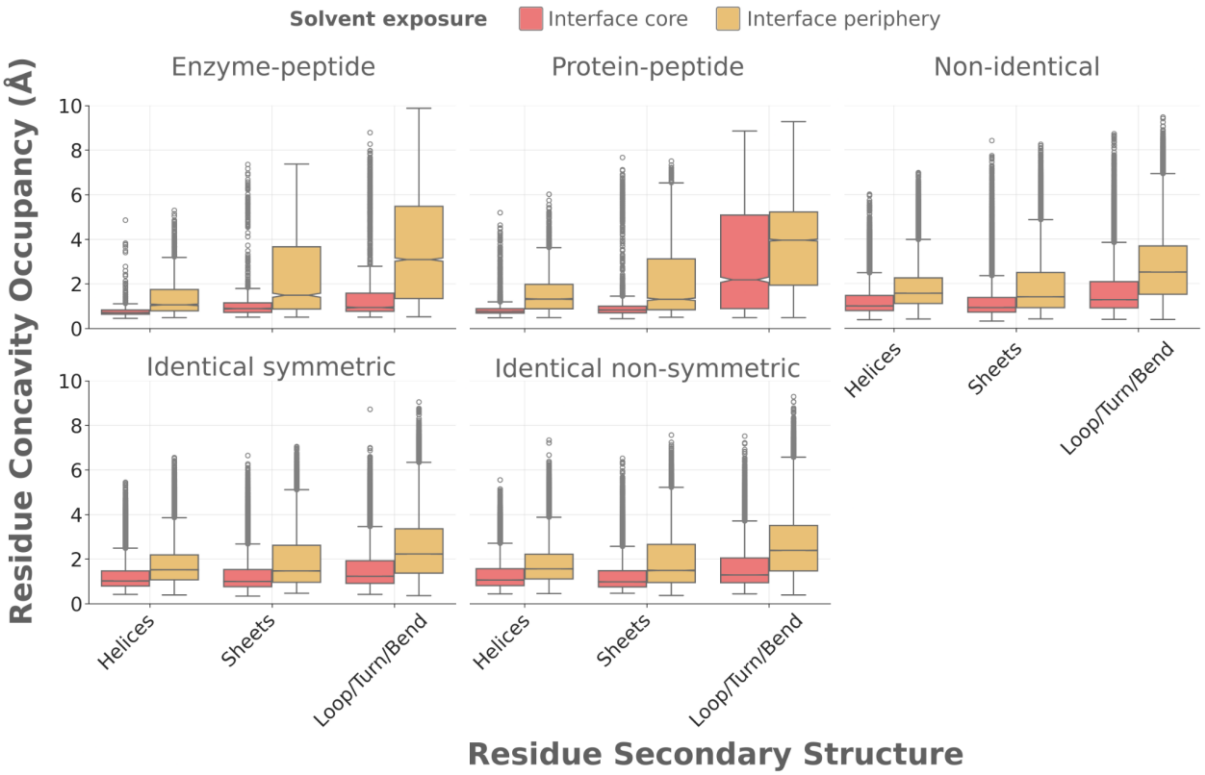


Figure S18 - Boxplot distributions of interface residue use of concavity by secondary structure, solvent exposure and interface type. Concavity is measured by Ghecom, representing the smallest spherical probe size that was able to enter a space around the partner protein's surface (where smaller values represent deeper binding).

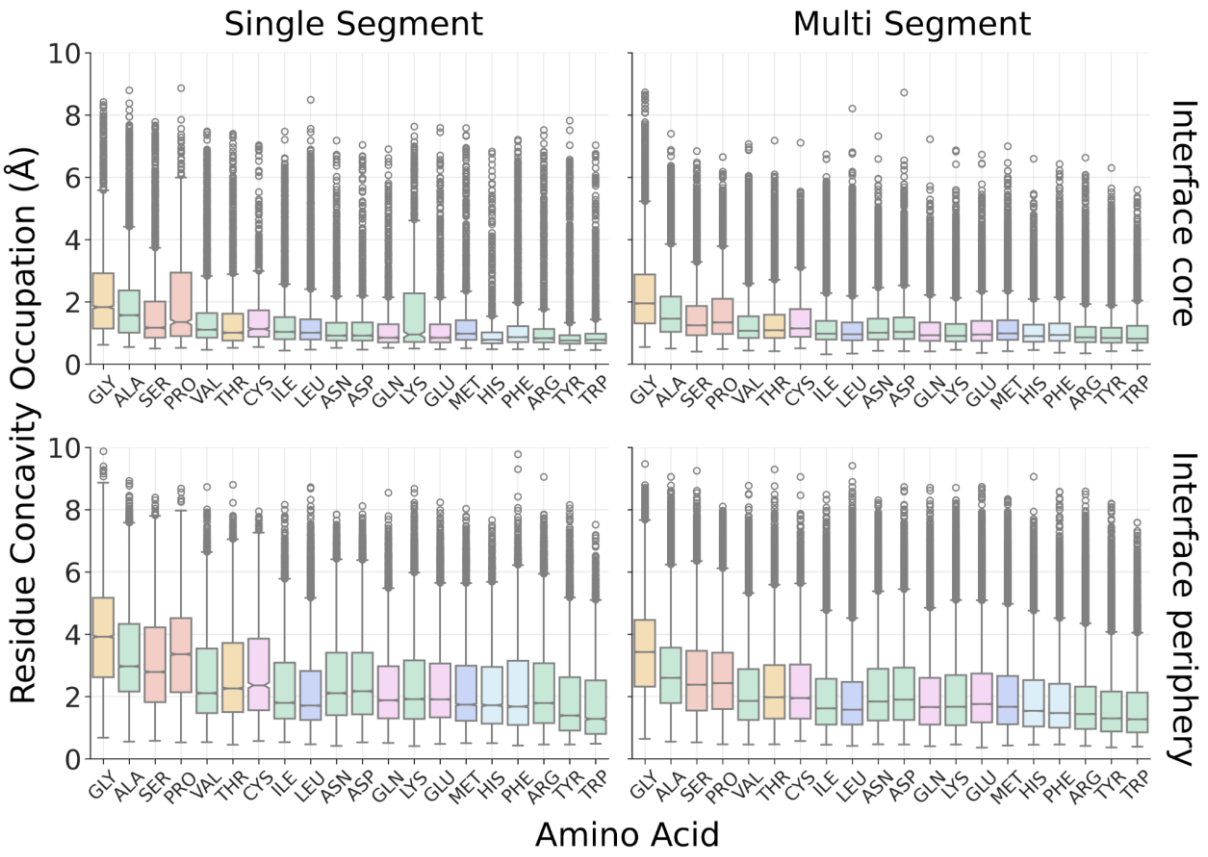


Figure S19 - Boxplot distributions of PPI residue use of concavity by amino acid. Residue distributions are coloured according to a modified version of the Lest colour scheme (Lesk, 2005) (small non-polar = orange, hydrophobic = green, polar = magenta, negatively charger = red, positively charged = blue). Concavity is measured by Ghecom, representing the smallest spherical probe size that was able to enter a space around the partner protein's surface (where smaller values represent deeper binding). Plots are divided by interface segmentation and solvent accessibility.

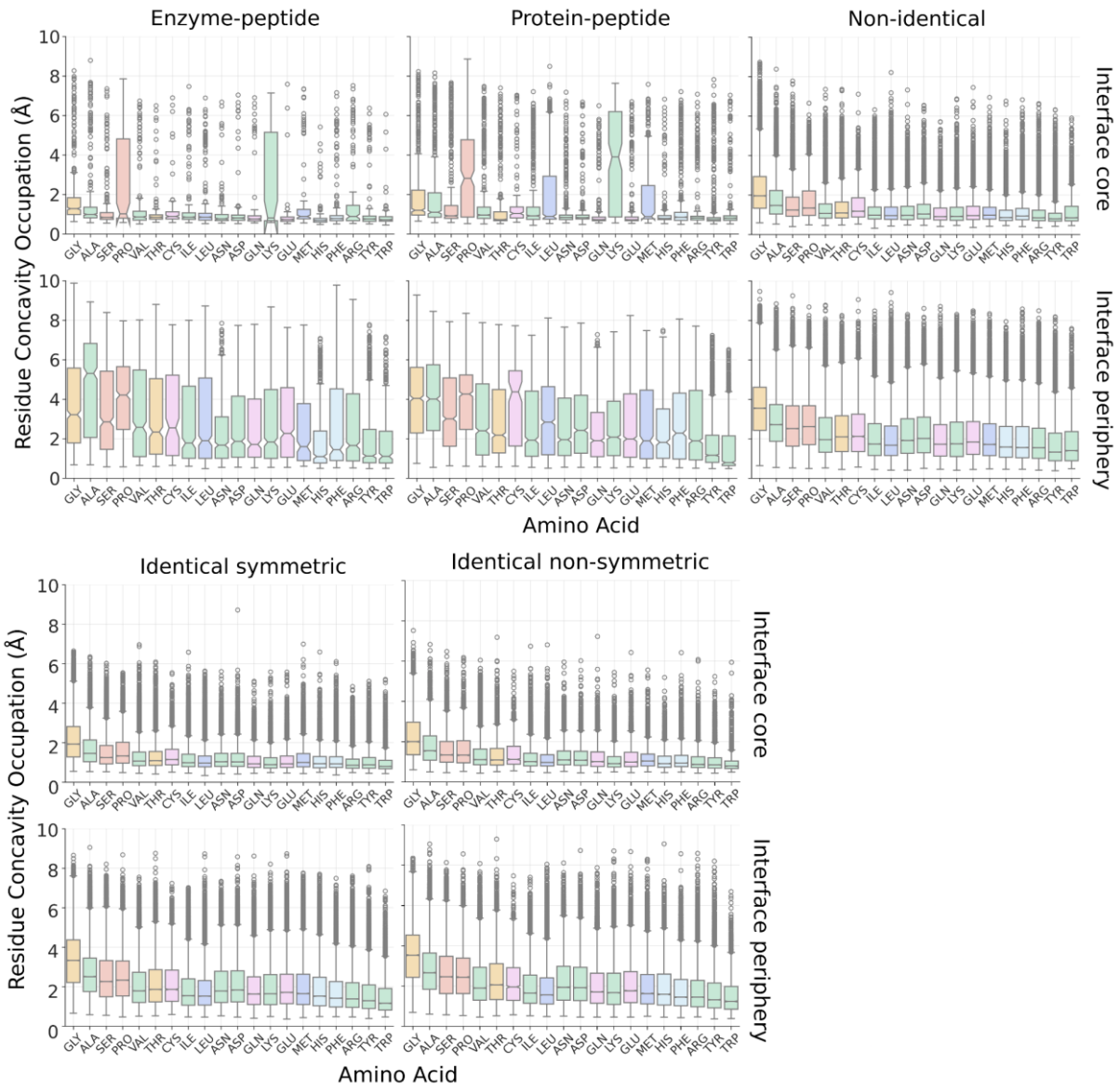


Figure S20 - Boxplot distributions of PPI residue use of concavity by amino acid. Residue distributions are coloured according to a modified version of the Lest colour scheme[1] (small non-polar = orange, hydrophobic = green, polar = magenta, negatively charger = red, positively charged = blue). Concavity is measured by Ghecom, representing the smallest spherical probe size that was able to enter a space around the partner protein's surface (where smaller values represent deeper binding). Plots are divided by interface type and solvent accessibility.

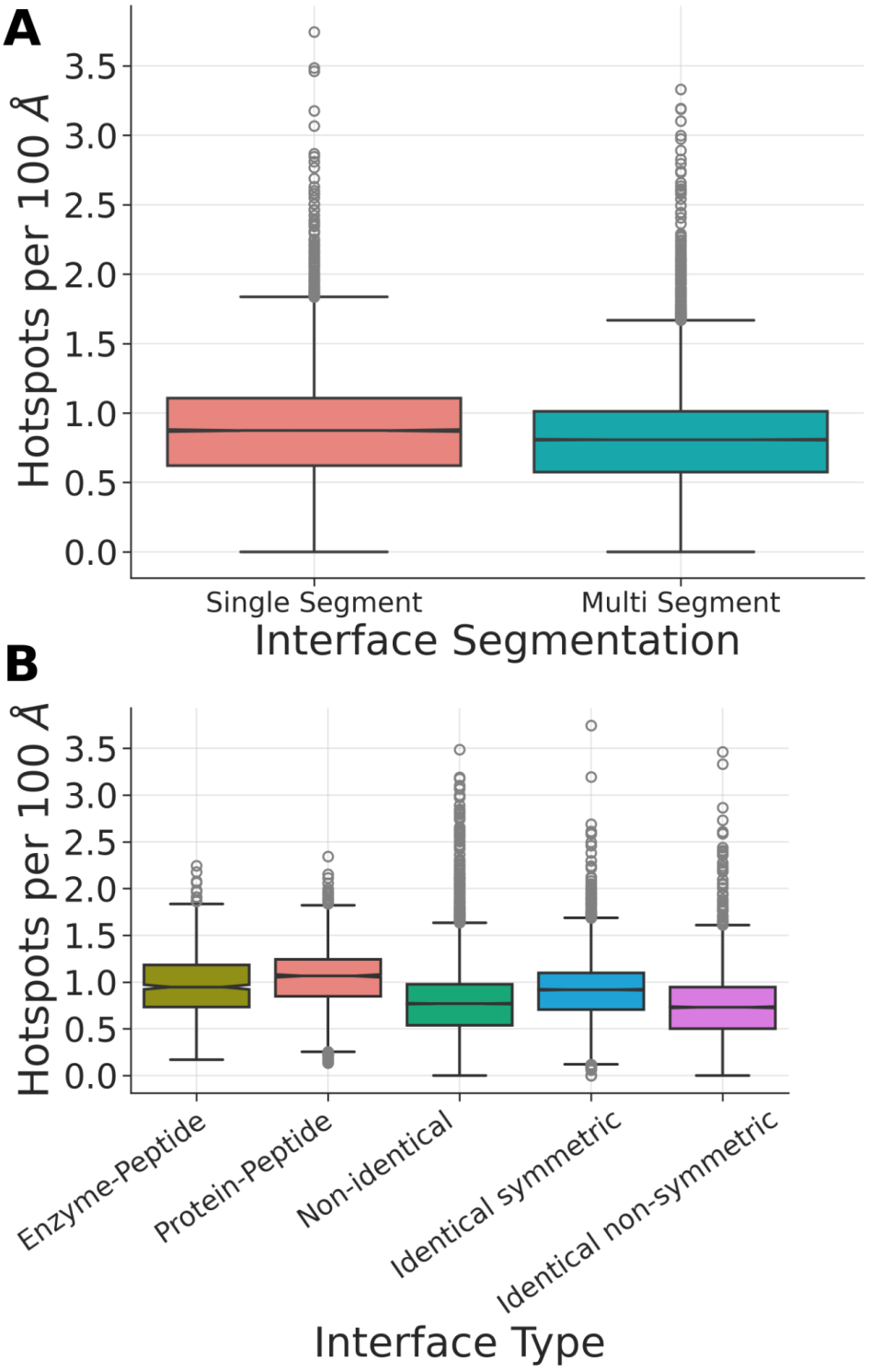


Figure S21 - Boxplot distributions of hotspots per 100Å² buried surface area. Distributions of mCSM-PPI predicted hotspots ($\Delta\Delta G^{Binding}$ > 1 kcal/mol) for PPI interfaces by A) interface segmentation and B) interface type.

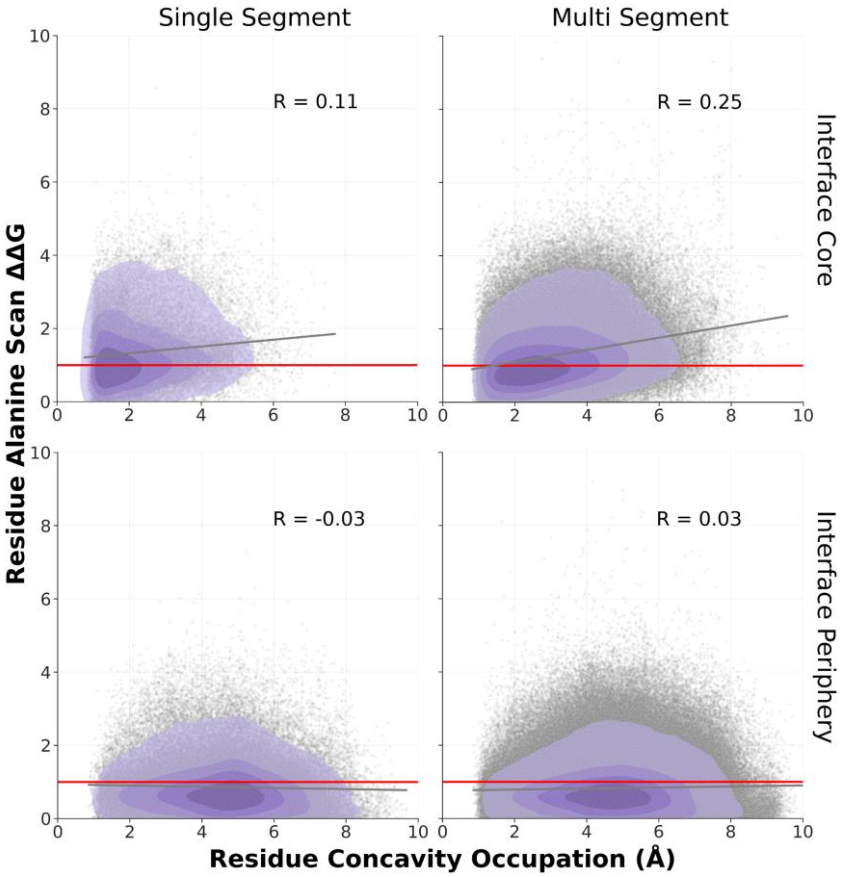


Figure S22 - Relationships between residue energetic hotspot predictions and use of concavity for interfaces by interface segmentation. Residue use of concavity as measured by Ghecom is shown on the abscissa, and mCSM-PPI alanine scanning $\Delta\Delta G^{Binding}$ predictions are shown on the ordinate. All residues originate from the deepest bound binding partner. The horizontal red lines show the threshold for mCSM-PPI predictions which consider a residue as a hotspot. Plots are divided by interface type. Linear model fitting is shown by gray lines. R values for Pearson correlation coefficient estimates are shown.

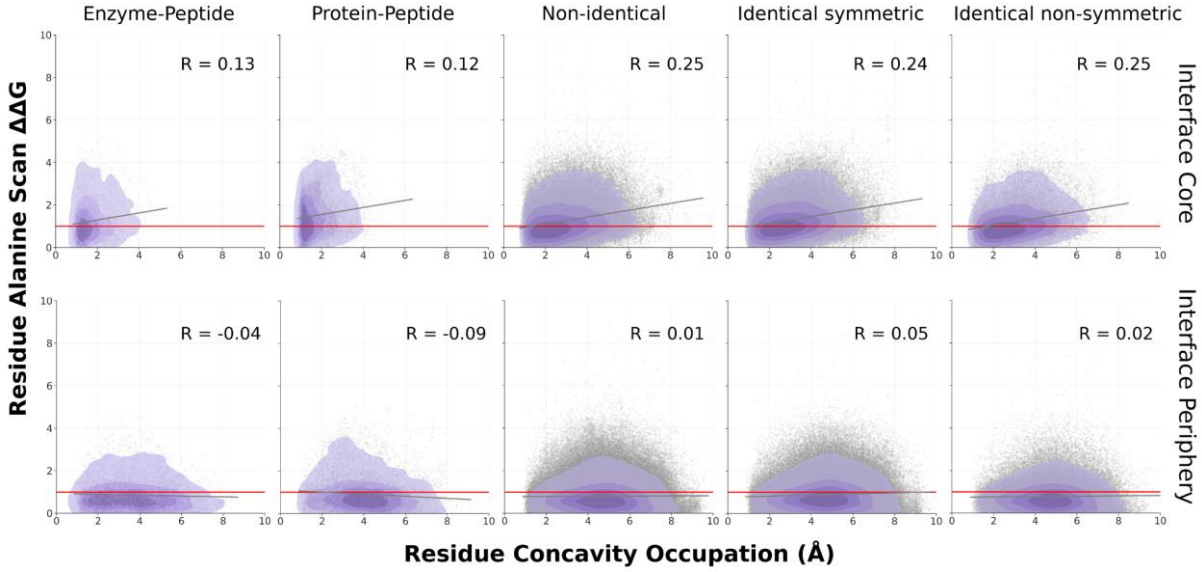


Figure S23 - Relationships between residue energetic hotspot predictions and use of concavity for different interfaces types. Residue use of concavity as measured by Ghecom is shown on the abscissa, and mCSM-PPI alanine scanning $\Delta\Delta G^{Binding}$ predictions are shown on the ordinate. All

residues originate from the deepest bound binding partner. The horizontal red lines show the threshold for mmCSM-PPI predictions to consider a residue as a hotspot. Plots are divided by interface classes. Linear model fitting is shown by gray lines. R values for linear Pearson correlation coefficient estimates are shown.

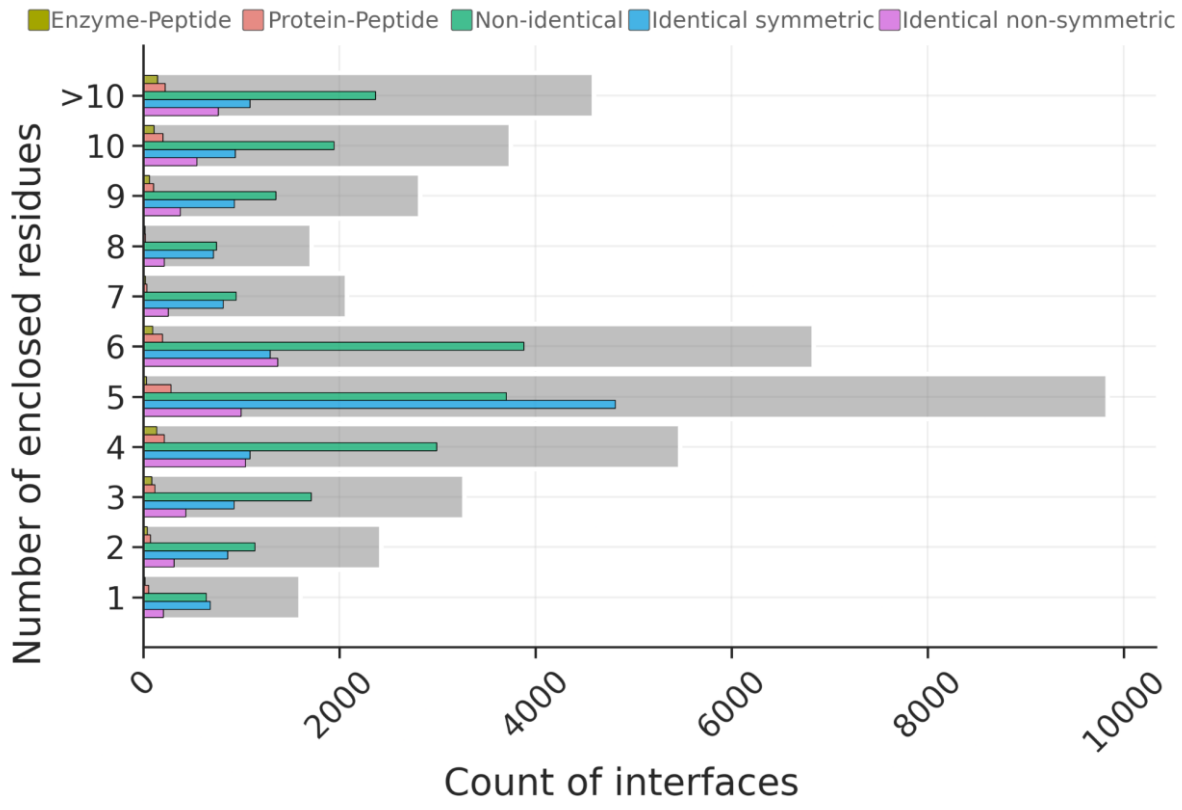


Figure S24 - Histogram distributions of numbers of deeply bound, solvent inaccessible residues in PPI interfaces. Distributions of the number of enclosed residues for the non-redundant set of PPI interfaces. Enclosed residues were solvent buried (interface core) residues with a concavity value of ≤ 4Å. The distribution for all interfaces is shown in gray, and the distributions for interface types in colours as per the legend.

Tables

Table S1 - Summary statistics and ANOVA analysis of PPI interface numbers of interface segmentation by interface type. ANOVA p-value < 0.05

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	1.01	0.13	822	1	1	3	1
Protein-peptide	a	1.03	0.24	1702	1	1	4	1
Non-identical	b	3.41	2.99	28165	1	3	32	1
Identical-symmetric	c	4.99	2.92	15920	1	4	33	4
Identical-nonsymmetric	d	4.3	3.95	8580	1	3	47	2

Table S2 - Summary statistics and ANOVA analysis of PPI interface planarity (Å) by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	4.54	1.82	10694	0.40	4.20	19.86	0.40
Multi	b	6.28	2.80	44415	0.51	5.53	20.03	0.51

Table S3 - Summary statistics and ANOVA analysis of PPI interface planarity (Å) by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	4.18	0.99	822	1.91	4.07	8.17	1.91
Protein-peptide	a	4.40	1.21	1702	0.70	4.32	9.51	0.69
Non-identical	b	5.74	2.65	28165	1.02	5.06	20.03	1.01
Identical-symmetric	c	6.52	2.72	15920	0.67	5.86	19.97	0.67
Identical-nonsymmetric	d	6.00	3.03	8580	0.40	5.12	19.97	0.40

Table S4 - Summary statistics and ANOVA analysis of proportions of secondary structure types in PPI interfaces in interface cores by interface type. Anova p-value < 0.05.

Segmentation	SST	Side	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	Helices	Smaller	a	0.34	0.38	10694	0.00	0.20	1.00	0.00
Single	Sheets	Smaller	b	0.08	0.17	10694	0.00	0.00	1.00	0.00
Single	Loops	Smaller	c	0.58	0.35	10694	0.00	0.60	1.00	1.00
Multi	Helices	Smaller	a	0.35	0.28	44415	0.00	0.32	1.00	0.00
Multi	Sheets	Smaller	d	0.16	0.20	10694	0.00	0.09	1.00	0.00
Multi	Loops	Smaller	e	0.49	0.23	10694	0.00	0.47	1.00	0.50
Single	Helices	Larger	f	0.41	0.34	10694	0.00	0.38	1.00	0.00
Single	Sheets	Larger	g	0.18	0.23	10694	0.00	0.07	1.00	0.00
Single	Loops	Larger	f	0.41	0.27	10694	0.00	0.38	1.00	0.00
Multi	Helices	Larger	a	0.35	0.27	44415	0.00	0.33	1.00	0.00
Multi	Sheets	Larger	d	0.18	0.20	44415	0.00	0.11	1.00	0.00
Multi	Loops	Larger	g	0.47	0.22	44415	0.00	0.46	1.00	0.50

Table S5 - Summary statistics and ANOVA analysis of proportions of secondary structure types in PPI interfaces in interface periphery by interface type. Anova p-value < 0.05.

Segmentation	SST	Side	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	Helices	Smaller	a	0.33	0.42	10694	0.00	0.00	1.00	0.00

Single	Sheets	Smaller	b	0.02	0.13	10694	0.00	0.00	1.00	0.00
Single	Loops	Smaller	a	0.34	0.42	10694	0.00	0.00	1.00	0.00
Multi	Helices	Smaller	c	0.36	0.32	44415	0.00	0.33	1.00	0.00
Multi	Sheets	Smaller	d	0.16	0.23	10694	0.00	0.00	1.00	0.00
Multi	Loops	Smaller	e	0.46	0.29	10694	0.00	0.46	1.00	0.50
Single	Helices	Larger	f	0.42	0.38	10694	0.00	0.40	1.00	0.00
Single	Sheets	Larger	dg	0.16	0.26	10694	0.00	0.00	1.00	0.00
Single	Loops	Larger	h	0.39	0.34	10694	0.00	0.33	1.00	0.00
Multi	Helices	Larger	i	0.37	0.32	44415	0.00	0.33	1.00	0.00
Multi	Sheets	Larger	g	0.17	0.24	44415	0.00	0.00	1.00	0.00
Multi	Loops	Larger	j	0.45	0.29	44415	0.00	0.43	1.00	0.00

Table S6 - Summary statistics and ANOVA analysis of PPI interface NIP by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.00078	0.00057	10694	0.00002	0.000628	0.00918	0.000265
Multi	b	0.0004	0.00041	44415	0	0.000256	0.01391	0.000081

Table S7 - Summary statistics and ANOVA analysis of PPI interface NIP by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.0008	0.00036	822	0.00018	0.000735	0.00237	0.000336
Protein-peptide	a	0.00078	0.00055	1702	0.00018	0.00064	0.00918	0.000242
Non-identical	b	0.00053	0.00048	28165	0	0.000373	0.00403	0.000009
Identical-symmetric	c	0.0003	0.00037	15920	0	0.000172	0.00443	0.000099
Identical-nonsymmetric	b	0.00052	0.00049	8580	0	0.00037	0.01391	0.000156

Table S8 - Summary statistics and ANOVA analysis of PPI interface NSc by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.00093	0.00063	10694	0	0.0007755	0.0071	0.000526
Multi	b	0.00051	0.00048	44415	0	0.000376	0.00864	0.000154

Table S9 - Summary statistics and ANOVA analysis of PPI interface NSc by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA	Mean	SD	N	Min	Med	Max	Mode

	Group							
Enzyme-peptide	a	0.00124	0.00056	822	0.00027	0.0011165	0.00396	0.000551
Protein-peptide	b	0.00104	0.00054	1702	0.00033	0.000902	0.00467	0.000558
Non-identical	c	0.00065	0.00055	28165	0	0.000504	0.00864	0.000154
Identical-symmetric	d	0.00038	0.00039	15920	0	0.000243	0.00701	0.000159
Identical-nonsymmetric	c	0.00065	0.00055	8580	0	0.000509	0.00572	0.000143

Table S10 - Summary statistics and ANOVA analysis of PPI interface BSA by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1173.17	957.48	10694	195.54	904.76	13496.83	386.22
Multi	b	2259.09	2259.79	44415	152.55	1621.27	27463.39	316.977

Table S11 - Summary statistics and ANOVA analysis of PPI interface BSA by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	828.08	294.95	822	315.57	772.13	2073.21	315.574
Protein-peptide	a	1005.93	455.06	1702	268.02	913.98	2403.32	268.024
Non-identical	b	1819.32	2053.82	28165	152.55	1253.41	22558.12	316.97
Identical-symmetric	c	2707.22	2030.73	15920	155.22	2230.63	26718.4	816.413
Identical-nonsymmetric	d	1900.42	2429.5	8580	199.08	1192.15	27463.39	318.71

Table S12 - Summary statistics and ANOVA analysis of PPI interface proportion of core residues by interface segmentation. ANOVA p-value < 0.05. Larger side refers to the side of the pairwise interface with fewer interacting residues than the smaller side.

Segmentation	Interface side	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	Larger	a	0.23	0.16	10694	0	0.19	0.94	0
Single	Smaller	b	0.1	0.14	10694	0	0.04	1	0
Multi	Larger	c	0.2	0.14	44415	0	0.21	0.62	0
Multi	Smaller	d	0.19	0.15	44415	0	0.19	0.77	0

Table S13 - Summary statistics and ANOVA analysis of PPI interface proportion of core residues by interface type. ANOVA p-value < 0.05. Larger side refers to the side of the pairwise interface with fewer interacting residues than the smaller side.

Interface type	Interface side	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	Larger	a	0.29	0.14	822	0.00	0.29	0.74	0.33

Enzyme-peptide	Smaller	b	0.09	0.14	822	0.00	0.00	0.67	0.00
Protein-peptide	Larger	a	0.30	0.18	1702	0.00	0.29	0.94	0.00
Protein-peptide	Smaller	c	0.15	0.19	1702	0.00	0.10	1.00	0.00
Non-identical	Larger	d	0.19	0.14	28124	0.00	0.18	0.64	0.00
Non-identical	Smaller	c	0.15	0.14	28124	0.00	0.13	0.82	0.00
Identical-symmetric	Larger	e	0.25	0.14	15912	0.00	0.27	0.62	0.00
Identical-symmetric	Smaller	e	0.26	0.14	15912	0.00	0.27	0.62	0.00
Identical-nonsymmetric	Larger	f	0.17	0.13	8549	0.00	0.17	0.62	0.00
Identical-nonsymmetric	Smaller	c	0.15	0.14	8549	0.00	0.14	0.61	0.00

Table S14 - Summary statistics and ANOVA analysis of VdW Clash interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.53	0.43	10694	-0.62	1.52	2.97	1.00
Multi	b	1.52	0.40	44415	-1.17	1.48	3.52	1.00

Table S15 - Summary statistics and ANOVA analysis of VdW Clash interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	1.57	0.32	822	0.71	1.57	2.51	0.70
Protein-peptide	b	1.47	0.30	1702	0.20	1.46	2.73	0.19
Non-identical	a	1.56	0.42	28165	-1.13	1.53	3.52	1.00
Identical-symmetric	c	1.42	0.37	15920	-1.17	1.39	2.97	1.00
Identical-nonsymmetric	a	1.58	0.43	8580	-1.16	1.57	2.96	1.00

Table S16 - Summary statistics and ANOVA analysis of VdW interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.27	0.43	10694	-0.82	1.26	2.76	1.00
Multi	b	1.26	0.39	44415	-1.31	1.22	3.03	0.00

Table S17 - Summary statistics and ANOVA analysis of VdW interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	1.29	0.33	822	0.41	1.29	2.29	0.40

Protein-peptide	b	1.18	0.30	1702	0.15	1.16	2.41	0.14
Non-identical	a	1.30	0.41	28165	-1.31	1.28	3.03	0.00
Identical-symmetric	b	1.16	0.37	15920	-1.07	1.13	2.76	0.00
Identical-nonsymmetric	c	1.32	0.42	8580	-1.24	1.30	2.61	0.00

Table S18 - Summary statistics and ANOVA analysis of Proximal interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	2.99	0.43	10694	1.00	2.99	4.38	1.00
Multi	a	2.99	0.38	44415	1.00	2.96	4.55	1.00

Table S19 - Summary statistics and ANOVA analysis of Proximal interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	3.06	0.31	822	2.22	3.07	3.99	2.21
Protein-peptide	b	2.96	0.29	1702	2.12	2.95	4.12	2.11
Non-identical	c	3.02	0.39	28165	1.00	3.01	4.52	1.00
Identical-symmetric	d	2.90	0.37	15920	1.00	2.88	4.55	1.00
Identical-nonsymmetric	a	3.05	0.42	8580	1.00	3.04	4.33	1.00

Table S20 - Summary statistics and ANOVA analysis of Hydrogen bonds at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.45	0.42	10694	-0.14	1.45	2.83	0.00
Multi	b	1.44	0.38	44415	0.00	1.41	2.96	0.00

Table S21 - Summary statistics and ANOVA analysis of Hydrogen bonds at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a,b	1.49	0.32	822	0.62	1.51	2.42	0.62
Protein-peptide	c	1.39	0.29	1702	0.20	1.38	2.56	0.19
Non-identical	b	1.47	0.40	28165	0.00	1.46	2.96	0.00
Identical-symmetric	d	1.36	0.36	15920	-0.14	1.33	2.96	0.00
Identical-nonsymmetric	a	1.50	0.42	8580	0.00	1.50	2.77	0.00

Table S22 - Summary statistics and ANOVA analysis of Weak Hydrogen bonds at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.30	0.42	10694	-0.88	1.30	2.69	0.00
Multi	a	1.29	0.38	44415	-1.02	1.27	2.84	0.00

Table S23 - Summary statistics and ANOVA analysis of Weak Hydrogen bonds at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	1.37	0.31	822	0.41	1.37	2.28	0.40
Protein-peptide	b	1.27	0.29	1702	0.26	1.27	2.38	0.25
Non-identical	c	1.31	0.39	28165	-1.02	1.30	2.82	0.00
Identical-symmetric	d	1.22	0.36	15920	-0.34	1.20	2.84	0.00
Identical-nonsymmetric	a	1.35	0.42	8580	-0.93	1.35	2.64	0.00

Table S24 - Summary statistics and ANOVA analysis of Ionic interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.40	0.50	10694	-1.84	0.43	2.09	0.00
Multi	b	0.44	0.45	44415	-1.77	0.44	2.31	0.00

Table S25 - Summary statistics and ANOVA analysis of Ionic interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.53	0.43	822	-0.99	0.57	1.62	0.00
Protein-peptide	b	0.45	0.35	1702	-1.14	0.50	1.73	0.00
Non-identical	c	0.41	0.45	28165	-1.84	0.41	1.99	0.00
Identical-symmetric	b	0.43	0.44	15920	-1.77	0.44	2.31	0.00
Identical-nonsymmetric	a	0.50	0.51	8580	-1.54	0.52	2.01	0.00

Table S26 - Summary statistics and ANOVA analysis of Hydrophobic interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.71	0.49	10694	-0.25	1.75	3.20	0.00
Multi	b	1.73	0.45	44415	-1.05	1.74	3.42	0.00

Table S27 - Summary statistics and ANOVA analysis of Hydrophobic interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	1.87	0.32	822	0.78	1.88	2.81	0.78
Protein-peptide	b	1.76	0.29	1702	0.69	1.77	2.88	0.68
Non-identical	c	1.72	0.48	28165	-1.05	1.75	3.27	0.00
Identical-symmetric	d	1.70	0.40	15920	-0.40	1.69	3.42	0.00
Identical-nonsymmetric	b	1.79	0.51	8580	-0.69	1.82	3.16	0.00

Table S28 - Summary statistics and ANOVA analysis of Carbonyl interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.44	0.54	10694	-1.65	0.48	2.07	0.00
Multi	b	0.48	0.45	44415	-1.66	0.46	2.53	0.00

Table S29 - Summary statistics and ANOVA analysis of Carbonyl interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.53	0.34	822	-0.63	0.54	1.36	-0.63
Protein-peptide	b	0.47	0.38	1702	-1.21	0.47	1.58	0.00
Non-identical	a	0.52	0.48	28165	-1.57	0.52	2.53	0.00
Identical-symmetric	c	0.35	0.44	15920	-1.66	0.35	2.10	0.00
Identical-nonsymmetric	a	0.54	0.48	8580	-1.50	0.54	2.20	0.00

Table S30 - Summary statistics and ANOVA analysis of Polar interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.63	0.42	10694	0.00	1.62	3.03	0.00
Multi	b	1.61	0.39	44415	0.00	1.58	3.17	0.00

Table S31 - Summary statistics and ANOVA analysis of Polar interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	ac	1.66	0.33	822	0.75	1.68	2.62	0.74
Protein-peptide	b	1.55	0.30	1702	0.39	1.54	2.77	0.39

Non-identical	c	1.64	0.41	28165	0.00	1.63	3.17	0.00
Identical-symmetric	b	1.54	0.37	15920	0.00	1.51	3.07	0.00
Identical-nonsymmetric	a	1.67	0.43	8580	0.00	1.67	2.98	0.00

Table S32 - Summary statistics and ANOVA analysis of Weak Polar interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	1.40	0.43	10694	-0.04	1.39	2.90	0.00
Multi	b	1.39	0.39	44415	0.00	1.35	3.23	0.00

Table S33 - Summary statistics and ANOVA analysis of Weak Polar interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	ac	1.43	0.32	822	0.61	1.45	2.40	0.61
Protein-peptide	b	1.34	0.29	1702	-0.04	1.34	2.53	-0.04
Non-identical	c	1.43	0.41	28165	0.00	1.42	3.23	0.00
Identical-symmetric	d	1.29	0.37	15920	0.00	1.27	2.90	0.00
Identical-nonsymmetric	e	1.44	0.42	8580	0.00	1.44	2.74	0.00

Table S34 - Summary statistics and ANOVA analysis of atom-ring Carbon-π interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.50	0.42	10694	-0.65	0.52	2.43	0.00
Multi	b	0.55	0.42	44415	-0.90	0.57	2.54	0.00

Table S35 - Summary statistics and ANOVA analysis of atom-ring Carbon-π interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.69	0.38	822	-0.19	0.77	1.99	0.00
Protein-peptide	a	0.73	0.37	1702	-0.17	0.80	1.86	0.00
Non-identical	b	0.51	0.42	28165	-0.77	0.52	2.34	0.00
Identical-symmetric	c	0.59	0.40	15920	-0.90	0.63	2.30	0.00
Identical-nonsymmetric	b	0.50	0.44	8580	-0.79	0.49	2.54	0.00

Table S36 - Summary statistics and ANOVA analysis of atom-ring Cation-π interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.03	0.18	10694	-1.00	0.00	1.45	0.00
Multi	b	0.00	0.23	44415	-1.33	0.00	1.43	0.00

Table S37 - Summary statistics and ANOVA analysis of atom-ring Cation-π interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.06	0.17	822	-0.30	0.00	0.98	0.00
Protein-peptide	a	0.07	0.21	1702	-0.33	0.00	1.31	0.00
Non-identical	b	0.01	0.21	28165	-1.28	0.00	1.43	0.00
Identical-symmetric	c	0.00	0.22	15920	-1.15	0.00	1.43	0.00
Identical-nonsymmetric	b	0.01	0.23	8580	-1.33	0.00	1.45	0.00

Table S38 - Summary statistics and ANOVA analysis of atom-ring Donor-π interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.09	0.25	10694	-1.12	0.00	1.62	0.00
Multi	b	0.07	0.29	44415	-1.26	0.00	1.72	0.00

Table S39 - Summary statistics and ANOVA analysis of atom-ring Donor-π interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.18	0.25	822	-0.20	0.00	1.60	0.00
Protein-peptide	a	0.17	0.26	1702	-0.32	0.02	1.35	0.00
Non-identical	b	0.07	0.28	28165	-1.26	0.00	1.72	0.00
Identical-symmetric	c	0.06	0.29	15920	-1.12	0.00	1.60	0.00
Identical-nonsymmetric	b	0.07	0.29	8580	-1.01	0.00	1.66	0.00

Table S40 - Summary statistics and ANOVA analysis of Ring-Ring interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.27	0.49	10694	-1.83	0.27	1.91	0.00
Multi	b	0.26	0.48	44415	-1.87	0.24	2.22	0.00

Table S41 - Summary statistics and ANOVA analysis of Ring-Ring interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.40	0.47	822	-0.75	0.43	1.52	0.00
Protein-peptide	b	0.33	0.39	1702	-0.99	0.42	1.46	0.00
Non-identical	c	0.28	0.47	28165	-1.57	0.27	1.91	0.00
Identical-symmetric	d	0.19	0.48	15920	-1.87	0.17	2.22	0.00
Identical-nonsymmetric	b	0.31	0.52	8580	-1.68	0.28	1.99	0.00

Table S42 - Summary statistics and ANOVA analysis of Amide-Amide interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.12	0.51	10694	-1.71	0.13	1.73	0.00
Multi	b	0.15	0.43	44415	-1.71	0.12	1.91	0.00

Table S43 - Summary statistics and ANOVA analysis of Amide-Amide interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.15	0.41	822	-1.05	0.20	1.29	0.00
Protein-peptide	b	0.00	0.41	1702	-1.26	0.00	1.27	0.00
Non-identical	c	0.20	0.45	28165	-1.53	0.19	1.91	0.00
Identical-symmetric	b	0.03	0.42	15920	-1.71	0.02	1.83	0.00
Identical-nonsymmetric	d	0.23	0.47	8580	-1.73	0.19	1.85	0.00

Table S44 - Summary statistics and ANOVA analysis of Amide-Ring interactions at PPI interfaces by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.74	0.52	10694	-1.12	0.79	2.28	0.00
Multi	b	0.75	0.47	44415	-1.04	0.75	2.60	0.00

Table S45 - Summary statistics and ANOVA analysis of Amide-Ring interactions at PPI interfaces by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.91	0.43	822	-0.15	0.95	1.89	0.00
Protein-peptide	b	0.84	0.40	1702	-0.18	0.88	1.86	0.00
Non-identical	c	0.77	0.48	28165	-0.92	0.79	2.29	0.00
Identical-symmetric	d	0.67	0.46	15920	-1.12	0.68	2.60	0.00

Identical-nonsymmetric	e	0.79	0.51	8580	-1.04	0.81	2.40	0.00

Table S46 - Summary statistics and ANOVA analysis of PPI interface use of concavity at their deepest and average depth by interface segmentation. ANOVA p-value < 0.05.

Segmentation	Concavity	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	average	a	4.37	1.34	10694	1.04	4.30	9.65	5.13
Single	deepest	b	0.81	0.84	10694	0.35	0.71	8.81	0.58
Multi	average	c	4.40	0.71	44415	1.77	4.27	9.13	4.10
Multi	deepest	d	1.03	0.54	44415	0.29	0.66	8.77	0.62

Table S47 - Summary statistics and ANOVA analysis of PPI interface use of concavity at their deepest and average depth by interface type. ANOVA p-value < 0.05.

Interface type	Concavity	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	average	a	3.11	1.21	822	1.04	2.98	9.65	1.04
Enzyme-peptide	deepest	b	0.75	0.78	822	0.45	0.61	7.76	0.58
Protein-peptide	average	c	3.45	1.45	1702	1.07	3.24	9.09	2.86
Protein-peptide	deepest	d	0.85	0.85	1702	0.44	0.61	6.78	0.56
Non-identical	average	e	4.54	0.91	28124	1.60	4.39	9.13	4.10
Non-identical	deepest	f	0.93	0.69	28124	0.29	0.69	8.81	0.62
Identical-symmetric	average	g	4.22	0.55	15912	2.86	4.13	8.07	4.32
Identical-symmetric	deepest	b	0.70	0.30	15912	0.34	0.63	7.35	0.60
Identical-nonsymmetric	average	h	4.58	0.74	8549	2.11	4.44	8.43	5.13
Identical-nonsymmetric	deepest	d	0.87	0.63	8549	0.36	0.69	7.77	0.63

Table S48 - Summary statistics and ANOVA analysis of interface length by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	10.99	0.65	10694	2	8	238	5
Multi	b	27.21	26.84	44415	3	20	516	5

Table S49 - Summary statistics and ANOVA analysis of interface length by interface segmentation. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	5.32	2.05	822	2	5	10	4
Protein-peptide	a	6.72	2.02	1702	2	7	10	9
Non-identical	b	20.69	23.34	28165	3	14	249	5

Identical-symmetric	c	33.41	24.23	15920	5	28	343	19
Identical-nonsymmetric	d	23.04	30.71	8580	3	14	516	5

Table S50 - Summary statistics and ANOVA analysis of chain length by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	107.54	164.42	10694	2	47	2136	9
Multi	b	284.27	236.85	44415	5	222	3795	213

Table S51 - Summary statistics and ANOVA analysis of chain length by interface segmentation. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	5.76	2.32	822	2	6	10	4
Protein-peptide	a	7.63	2.07	1702	2	8	10	9
Non-identical	b	246.89	263.23	28165	11	178	3795	11
Identical-symmetric	c	277.15	173.05	15920	32	250	2750	141
Identical-nonsymmetric	c	281.32	230.72	8580	11	229	3303	121

Table S52 - Summary statistics and ANOVA analysis of interface residue use of concavity, by interface segmentation, secondary structure (SST), and solvent accessibility. ANOVA p-value < 0.05.

Segmentation	Exposure	SST	ANOVA group	Mean	SD	N	Min	Med	Max	Mode
Single	core	helices	ag	1.23	0.67	30698	0.45	0.98	5.55	0.75
Single	core	sheets	b	1.29	1.07	13539	0.44	0.91	8.42	0.65
Single	core	loops	c	2.10	1.68	14655	0.47	1.30	8.86	0.83
Single	periphery	helices	d	1.80	0.81	97394	0.41	1.64	6.62	1.43
Single	periphery	sheets	e	2.38	1.62	26462	0.43	1.81	8.25	0.78
Single	periphery	loops	f	3.30	1.67	107417	0.42	3.22	9.88	1.33
Multi	core	helices	a	1.24	0.65	272412	0.38	1.01	6.01	0.74
Multi	core	sheets	g	1.25	0.76	158120	0.32	0.97	7.61	0.68
Multi	core	loops	h	1.61	0.96	211237	0.40	1.26	8.73	0.87
Multi	periphery	helices	I	1.75	0.88	601464	0.39	1.53	7.34	0.88
Multi	periphery	sheets	d	1.80	1.09	254394	0.37	1.42	8.18	0.72
Multi	periphery	loops	j	2.57	1.35	898933	0.36	2.34	9.47	1.03

Table S53 - Summary statistics and ANOVA analysis of interface residue use of concavity, by interface type, secondary structure (SST), and solvent accessibility. ANOVA p-value < 0.05.

Interface type	Exposure	SST	ANOVA group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	core	helices	a	0.79	0.35	1003	0.45	0.71	4.86	0.66
Enzyme-peptide	core	sheets	b,g,p,r	1.23	1.11	1294	0.51	0.89	7.36	0.65
Enzyme-peptide	core	loops	c,n,o,s	1.86	1.90	1683	0.51	0.94	8.79	0.82
Enzyme-peptide	periphery	helices	d	1.40	0.88	1849	0.49	1.06	5.30	0.73
Enzyme-peptide	periphery	sheets	e	2.41	1.91	2287	0.51	1.49	7.37	0.80
Enzyme-peptide	periphery	loops	f	3.49	2.22	7790	0.52	3.09	9.88	0.97
Protein-peptide	core	helices	a	0.87	0.41	4554	0.48	0.77	5.20	0.71
Protein-peptide	core	sheets	g	1.11	1.04	3383	0.44	0.83	7.67	0.74
Protein-peptide	core	loops	h	3.05	2.27	3014	0.49	2.19	8.86	0.78
Protein-peptide	periphery	helices	l	1.54	0.83	7565	0.49	1.32	6.02	0.82
Protein-peptide	periphery	sheets	j	2.18	1.75	4091	0.50	1.31	7.51	0.74
Protein-peptide	periphery	loops	k	3.75	1.92	14942	0.49	3.96	9.28	1.02
Non-identical	core	helices	b	1.24	0.66	117380	0.38	1.00	6.01	0.76
Non-identical	core	sheets	b	1.22	0.78	74336	0.32	0.93	8.42	0.68
Non-identical	core	loops	l	1.66	1.03	98715	0.40	1.28	8.73	0.87
Non-identical	periphery	helices	m	1.79	0.90	307932	0.41	1.57	6.98	0.98
Non-identical	periphery	sheets	n	1.82	1.15	126756	0.42	1.41	8.25	0.71
Non-identical	periphery	loops	o	2.72	1.42	496872	0.39	2.52	9.47	1.08
Identical-symmetric	core	helices	p	1.24	0.63	146895	0.42	1.02	5.45	0.78
Identical-symmetric	core	sheets	b	1.30	0.78	69368	0.34	1.00	6.65	0.69
Identical-symmetric	core	loops	l	1.57	0.93	92179	0.42	1.23	8.72	0.86
Identical-symmetric	periphery	helices	q	1.72	0.84	271587	0.39	1.52	6.56	0.92
Identical-symmetric	periphery	sheets	c	1.87	1.13	102161	0.47	1.47	7.05	0.72
Identical-symmetric	periphery	loops	e	2.48	1.32	331673	0.36	2.23	9.05	1.03
Identical-nonsymmetric	core	helices	r	1.30	0.69	33278	0.44	1.06	5.55	0.73
Identical-nonsymmetric	core	sheets	b	1.28	0.80	23278	0.47	0.98	6.52	0.66
Identical-nonsymmetric	core	loops	l	1.64	0.97	30301	0.44	1.29	7.52	0.84
Identical-nonsymmetric	periphery	helices	c	1.77	0.86	109925	0.45	1.56	7.34	1.38
Identical-nonsymmetric	periphery	sheets	s	1.88	1.14	45561	0.37	1.49	7.57	0.68
Identical-nonsymmetric	periphery	loops	t	2.60	1.35	155073	0.39	2.39	9.29	1.08

Table S54 - Summary statistics and ANOVA analysis of interface residue use of concavity, by amino acid, and by interface segmentation. ANOVA p-value < 0.05.

Segmentation	Amino Acid	Exposure	ANOVA group	Mean	SD	N	Min	Med	Max	Mode
Single	GLY	core	klmn	2.26	1.47	3631	0.62	1.83	8.42	0.93
Single	ALA	core	q	1.89	1.21	5415	0.55	1.57	8.79	0.96
Single	SER	core	rs	1.68	1.29	3174	0.50	1.17	7.78	0.84
Single	PRO	core	op	2.07	1.55	1521	0.52	1.35	8.86	0.87
Single	VAL	core	tu	1.50	1.11	5308	0.46	1.11	7.48	0.79
Single	THR	core	uv	1.42	1.06	2719	0.52	1.01	7.40	0.71
Single	CYS	core	tu	1.50	1.04	1050	0.55	1.13	7.03	0.88
Single	ILE	core	uv	1.38	1.00	5058	0.44	1.04	7.47	0.84
Single	LEU	core	uv	1.38	1.06	8873	0.47	1.01	8.49	0.80
Single	ASN	core	xy	1.23	0.86	2165	0.52	0.92	7.18	0.71
Single	ASP	core	wxy	1.24	0.91	1561	0.47	0.92	7.04	0.72
Single	GLN	core	xyz	1.21	0.96	1480	0.52	0.85	6.91	0.66
Single	LYS	core	pq	1.93	1.87	735	0.50	0.95	7.63	0.70
Single	GLU	core	xyzA	1.19	0.94	1685	0.48	0.85	7.59	0.68
Single	MET	core	vw	1.35	1.06	1959	0.51	0.98	7.58	0.70
Single	HIS	core	B	1.06	0.85	1256	0.48	0.78	6.83	0.61
Single	PHE	core	xyz	1.21	0.97	4040	0.48	0.87	7.21	0.73
Single	ARG	core	vwx	1.29	1.21	1510	0.48	0.83	7.52	0.74
Single	TYR	core	B	1.01	0.87	3915	0.45	0.76	7.82	0.64
Single	TRP	core	B	1.04	0.86	1837	0.45	0.78	7.03	0.67
Multi	GLY	core	mn	2.22	1.14	46276	0.55	1.95	8.73	1.15
Multi	ALA	core	r	1.71	0.87	55637	0.50	1.46	7.40	0.97
Multi	SER	core	t	1.52	0.80	39923	0.40	1.25	6.84	0.89
Multi	PRO	core	s	1.63	0.87	26302	0.49	1.34	6.65	0.90
Multi	VAL	core	vw	1.33	0.74	55805	0.44	1.07	7.07	0.80
Multi	THR	core	v	1.35	0.75	35527	0.42	1.09	7.18	0.78
Multi	CYS	core	u	1.43	0.77	9881	0.51	1.15	7.11	0.93
Multi	ILE	core	xy	1.22	0.67	50814	0.32	0.98	6.74	0.79
Multi	LEU	core	xyz	1.19	0.66	85502	0.34	0.96	8.21	0.73
Multi	ASN	core	wx	1.26	0.71	20223	0.43	1.01	7.32	0.74
Multi	ASP	core	wx	1.27	0.69	20158	0.42	1.04	8.72	0.78

Multi	GLN	core	yzA	1.16	0.62	18732	0.41	0.93	7.22	0.77
Multi	LYS	core	xyzA	1.17	0.70	10210	0.46	0.91	6.87	0.70
Multi	GLU	core	xyzA	1.18	0.64	20668	0.36	0.95	6.73	0.70
Multi	MET	core	xy	1.23	0.69	22999	0.42	0.99	7.00	0.73
Multi	HIS	core	zA	1.12	0.61	14682	0.45	0.90	6.60	0.71
Multi	PHE	core	yzA	1.15	0.63	41058	0.37	0.94	6.43	0.76
Multi	ARG	core	AB	1.08	0.60	21951	0.35	0.86	6.63	0.67
Multi	TYR	core	B	1.06	0.62	32928	0.42	0.84	6.30	0.68
Multi	TRP	core	B	1.06	0.60	12493	0.44	0.82	5.60	0.69
Single	GLY	periphery	a	3.95	1.68	10820	0.68	3.92	9.88	4.73
Single	ALA	periphery	c	3.35	1.62	12685	0.55	2.97	8.93	2.31
Single	SER	periphery	d	3.12	1.59	13044	0.58	2.79	8.39	1.33
Single	PRO	periphery	c	3.39	1.54	10122	0.53	3.36	8.68	1.51
Single	VAL	periphery	g	2.61	1.50	12089	0.54	2.11	8.73	1.66
Single	THR	periphery	f	2.71	1.53	12299	0.45	2.26	8.80	1.88
Single	CYS	periphery	e	2.84	1.62	1692	0.57	2.36	7.95	1.79
Single	ILE	periphery	jk	2.31	1.40	11545	0.54	1.80	8.16	1.52
Single	LEU	periphery	no	2.19	1.33	21408	0.47	1.71	8.72	1.35
Single	ASN	periphery	h	2.52	1.42	10873	0.42	2.11	7.85	1.85
Single	ASP	periphery	gh	2.56	1.44	13299	0.53	2.17	8.12	1.35
Single	GLN	periphery	jkl	2.29	1.34	11699	0.51	1.88	8.55	1.34
Single	LYS	periphery	i	2.37	1.43	16946	0.41	1.92	8.67	1.08
Single	GLU	periphery	ij	2.34	1.36	18572	0.48	1.91	8.24	1.50
Single	MET	periphery	jklm	2.28	1.45	5919	0.50	1.74	8.03	1.42
Single	HIS	periphery	mno	2.20	1.40	5819	0.50	1.72	7.67	1.35
Single	PHE	periphery	lmn	2.23	1.47	9321	0.43	1.68	9.78	1.06
Single	ARG	periphery	klm	2.27	1.45	17946	0.46	1.79	9.05	0.73
Single	TYR	periphery	p	1.96	1.40	10926	0.46	1.39	8.15	0.83
Single	TRP	periphery	q	1.84	1.35	4249	0.49	1.28	7.52	0.64
Multi	GLY	periphery	b	3.44	1.40	96823	0.64	3.43	9.47	3.43
Multi	ALA	periphery	e	2.78	1.27	91491	0.55	2.60	9.05	1.83
Multi	SER	periphery	g	2.61	1.29	106607	0.53	2.38	9.25	1.53
Multi	PRO	periphery	g	2.60	1.24	89625	0.47	2.43	8.10	1.56
Multi	VAL	periphery	o	2.18	1.18	80025	0.46	1.86	8.77	1.04

Multi	THR	periphery	lmn	2.25	1.19	96688	0.47	1.98	9.29	1.02
Multi	CYS	periphery	jkl	2.29	1.27	12347	0.57	1.95	9.05	1.04
Multi	ILE	periphery	p	1.95	1.10	68657	0.45	1.62	8.48	0.99
Multi	LEU	periphery	q	1.91	1.07	125722	0.42	1.58	9.41	0.98
Multi	ASN	periphery	o	2.15	1.16	89736	0.47	1.84	8.31	1.28
Multi	ASP	periphery	no	2.20	1.18	116906	0.46	1.90	8.73	1.12
Multi	GLN	periphery	p	1.96	1.09	91517	0.40	1.66	8.71	0.80
Multi	LYS	periphery	p	1.99	1.14	128948	0.46	1.67	8.71	0.76
Multi	GLU	periphery	p	2.05	1.13	146375	0.36	1.76	8.74	0.86
Multi	MET	periphery	p	2.04	1.24	37029	0.43	1.67	8.34	0.91
Multi	HIS	periphery	q	1.90	1.11	51098	0.45	1.54	9.06	1.12
Multi	PHE	periphery	q	1.82	1.08	63823	0.46	1.47	8.58	0.96
Multi	ARG	periphery	q	1.76	1.04	156945	0.42	1.44	8.59	0.79
Multi	TYR	periphery	s	1.66	1.04	78700	0.37	1.30	8.20	0.81
Multi	TRP	periphery	s	1.62	1.00	25729	0.39	1.27	7.59	0.70

Table S55 - Summary statistics and ANOVA analysis of interface residue use of concavity, by amino acid, and by interface type. ANOVA p-value < 0.05.

Interface type	Amino Acid	Exposure	ANOVA group	Mean	SD	N	Min	Med	Max	Mode
Enzyme peptide	GLY	core	nop	1.80	1.51	423	0.63	1.28	8.27	0.89
Enzyme peptide	ALA	core	opq	1.66	1.75	343	0.57	0.99	8.79	0.97
Enzyme peptide	SER	core	qrs	1.34	1.35	280	0.55	0.84	7.36	0.80
Enzyme peptide	PRO	core	ij	2.51	2.22	134	0.58	1.02	7.85	0.66
Enzyme peptide	VAL	core	pqr	1.43	1.50	239	0.51	0.86	6.74	0.68
Enzyme peptide	THR	core	rst	1.18	1.10	141	0.59	0.85	6.50	0.71
Enzyme peptide	CYS	core	qrst	1.25	1.19	108	0.58	0.89	6.90	0.69
Enzyme peptide	ILE	core	qrst	1.28	1.30	227	0.57	0.85	7.47	0.82
Enzyme peptide	LEU	core	qrs	1.33	1.35	322	0.52	0.85	6.89	0.84
Enzyme peptide	ASN	core	rst	1.15	1.13	127	0.55	0.77	6.68	0.66
Enzyme peptide	ASP	core	rst	1.12	1.20	161	0.52	0.81	7.04	0.68
Enzyme peptide	GLN	core	qrst	1.29	1.53	112	0.56	0.76	6.91	0.65
Enzyme peptide	LYS	core	hij	2.60	2.41	65	0.58	0.81	7.14	0.59
Enzyme peptide	GLU	core	t	0.92	0.95	108	0.52	0.73	7.59	0.64
Enzyme peptide	MET	core	nopq	1.74	1.86	107	0.56	0.89	7.35	0.78

Enzyme peptide	HIS	core	t	0.81	0.59	264	0.48	0.69	5.42	0.61
Enzyme peptide	PHE	core	rst	1.16	1.23	307	0.51	0.78	7.16	0.67
Enzyme peptide	ARG	core	jklmnop	1.96	2.11	107	0.53	0.89	7.52	0.62
Enzyme peptide	TYR	core	st	1.09	1.15	246	0.51	0.76	6.37	0.62
Enzyme peptide	TRP	core	t	0.87	0.65	159	0.45	0.72	6.06	0.65
Protein peptide	GLY	core	jkl	2.09	1.87	555	0.65	1.20	8.23	0.93
Protein peptide	ALA	core	jk	2.12	2.06	579	0.57	1.10	8.16	0.83
Protein peptide	SER	core	nopq	1.72	1.82	659	0.58	0.92	7.67	0.80
Protein peptide	PRO	core	fgh	2.91	2.08	240	0.52	2.82	8.86	0.66
Protein peptide	VAL	core	mnop	1.83	1.86	700	0.51	0.96	7.48	0.70
Protein peptide	THR	core	qrst	1.28	1.35	630	0.52	0.73	7.40	0.67
Protein peptide	CYS	core	opq	1.68	1.68	150	0.59	1.04	7.03	1.07
Protein peptide	ILE	core	nopq	1.73	1.71	612	0.44	0.92	7.21	0.71
Protein peptide	LEU	core	jklmn	1.99	1.93	1134	0.53	0.89	8.49	0.73
Protein peptide	ASN	core	rst	1.12	0.98	698	0.56	0.84	7.18	0.73
Protein peptide	ASP	core	rst	1.15	1.08	376	0.48	0.83	6.68	0.82
Protein peptide	GLN	core	rst	1.19	1.23	330	0.56	0.73	6.50	0.71
Protein peptide	LYS	core	cd	3.59	2.53	154	0.56	3.90	7.63	6.25
Protein peptide	GLU	core	rst	1.13	1.24	458	0.54	0.72	6.70	0.68
Protein peptide	MET	core	jklmno	1.96	1.97	206	0.55	0.89	7.58	0.82
Protein peptide	HIS	core	st	1.09	1.03	321	0.52	0.79	6.83	0.75
Protein peptide	PHE	core	pq	1.49	1.57	595	0.49	0.85	7.21	0.65
Protein peptide	ARG	core	qr	1.42	1.48	389	0.52	0.81	7.08	0.86
Protein peptide	TYR	core	t	0.97	1.01	1508	0.51	0.73	7.82	0.68
Protein peptide	TRP	core	t	1.03	0.97	657	0.51	0.80	7.03	0.80
Non-identical	GLY	core	j	2.28	1.22	19594	0.59	1.96	8.73	1.15
Non-identical	ALA	core	nop	1.75	0.94	23061	0.53	1.47	8.38	0.90
Non-identical	SER	core	pq	1.53	0.84	18854	0.40	1.24	7.78	0.97
Non-identical	PRO	core	opq	1.67	0.91	12201	0.49	1.34	6.65	0.87
Non-identical	VAL	core	qr	1.34	0.78	24497	0.46	1.07	7.36	0.82
Non-identical	THR	core	qr	1.37	0.78	15145	0.48	1.09	7.36	0.79
Non-identical	CYS	core	pq	1.48	0.82	5007	0.53	1.18	7.11	0.88
Non-identical	ILE	core	rst	1.23	0.71	22447	0.32	0.98	6.32	0.79
Non-identical	LEU	core	rst	1.23	0.74	38787	0.43	0.96	8.21	0.71

Non-identical	ASN	core	qrst	1.25	0.75	9335	0.48	0.97	7.32	0.74
Non-identical	ASP	core	qrs	1.29	0.73	9140	0.47	1.04	6.55	0.78
Non-identical	GLN	core	rst	1.15	0.63	8826	0.41	0.91	5.72	0.77
Non-identical	LYS	core	qrst	1.24	0.81	4847	0.46	0.92	6.87	0.71
Non-identical	GLU	core	rst	1.21	0.68	9135	0.36	0.96	7.45	0.69
Non-identical	MET	core	rst	1.23	0.72	10406	0.42	0.98	6.95	0.73
Non-identical	HIS	core	st	1.10	0.63	6296	0.45	0.86	6.76	0.66
Non-identical	PHE	core	rst	1.17	0.66	19772	0.40	0.94	6.83	0.71
Non-identical	ARG	core	rst	1.10	0.67	10243	0.35	0.85	6.63	0.67
Non-identical	TYR	core	t	1.03	0.65	16407	0.42	0.80	6.32	0.68
Non-identical	TRP	core	rst	1.12	0.65	6431	0.45	0.85	5.90	0.67
Identical symmetric	GLY	core	j	2.16	1.08	22232	0.55	1.93	6.65	0.98
Identical symmetric	ALA	core	opq	1.69	0.83	29322	0.54	1.46	6.36	0.94
Identical symmetric	SER	core	pq	1.50	0.79	17652	0.53	1.25	6.03	0.78
Identical symmetric	PRO	core	opq	1.60	0.86	11538	0.49	1.33	5.55	1.08
Identical symmetric	VAL	core	qrs	1.32	0.71	28250	0.45	1.07	6.97	0.80
Identical symmetric	THR	core	qrs	1.34	0.74	17211	0.42	1.09	6.10	0.87
Identical symmetric	CYS	core	qr	1.39	0.72	4468	0.51	1.15	5.55	0.82
Identical symmetric	ILE	core	rst	1.23	0.66	25860	0.46	0.99	6.59	0.80
Identical symmetric	LEU	core	rst	1.17	0.61	44069	0.34	0.97	5.63	0.74
Identical symmetric	ASN	core	qrst	1.26	0.68	9124	0.43	1.04	5.61	0.73
Identical symmetric	ASP	core	qrst	1.25	0.66	8834	0.42	1.03	8.72	0.79
Identical symmetric	GLN	core	rst	1.14	0.59	8175	0.45	0.93	5.12	0.70
Identical symmetric	LYS	core	rst	1.11	0.63	4173	0.49	0.89	5.59	0.70
Identical symmetric	GLU	core	rst	1.14	0.58	9663	0.47	0.93	5.58	0.69
Identical symmetric	MET	core	rst	1.24	0.68	11220	0.44	1.00	7.00	0.76
Identical symmetric	HIS	core	rst	1.13	0.61	7260	0.50	0.94	6.60	0.71

Identical symmetric	PHE	core	rst	1.14	0.62	19627	0.37	0.93	6.11	0.76
Identical symmetric	ARG	core	st	1.05	0.55	9601	0.49	0.86	4.85	0.67
Identical symmetric	TYR	core	st	1.09	0.61	14506	0.45	0.88	5.12	0.70
Identical symmetric	TRP	core	t	1.01	0.58	5657	0.44	0.80	5.21	0.67
Identical non-symmetric	GLY	core	j	2.27	1.13	7103	0.61	2.00	7.52	1.15
Identical non-symmetric	ALA	core	nop	1.79	0.90	7747	0.50	1.55	6.83	1.01
Identical non-symmetric	SER	core	opq	1.62	0.86	5652	0.46	1.34	6.48	1.01
Identical non-symmetric	PRO	core	opq	1.62	0.88	3710	0.52	1.33	6.18	0.75
Identical non-symmetric	VAL	core	qr	1.40	0.81	7427	0.44	1.11	6.05	0.77
Identical non-symmetric	THR	core	qr	1.37	0.76	5119	0.51	1.09	7.18	0.82
Identical non-symmetric	CYS	core	qr	1.42	0.78	1198	0.55	1.12	5.50	0.73
Identical non-symmetric	ILE	core	qrst	1.25	0.70	6726	0.50	1.01	6.74	0.74
Identical non-symmetric	LEU	core	rst	1.20	0.64	10063	0.46	0.97	6.81	0.84
Identical non-symmetric	ASN	core	qrs	1.31	0.68	3104	0.52	1.10	5.96	0.92
Identical non-symmetric	ASP	core	qrs	1.30	0.68	3208	0.45	1.08	6.02	0.85
Identical non-symmetric	GLN	core	qrst	1.25	0.70	2769	0.47	1.01	7.22	0.75
Identical non-symmetric	LYS	core	rst	1.13	0.65	1706	0.50	0.91	5.44	0.70
Identical non-symmetric	GLU	core	qrst	1.24	0.71	2989	0.46	0.99	6.43	0.70
Identical non-symmetric	MET	core	rst	1.23	0.65	3019	0.50	1.04	5.56	0.73
Identical non-symmetric	HIS	core	rst	1.14	0.63	1797	0.47	0.91	5.19	0.62
Identical non-symmetric	PHE	core	rst	1.17	0.67	4797	0.49	0.94	6.42	0.75
Identical non-symmetric	ARG	core	rst	1.11	0.63	3121	0.44	0.87	6.08	0.66
Identical non-symmetric	TYR	core	st	1.07	0.60	4176	0.44	0.85	5.24	0.66
Identical non-symmetric	TRP	core	t	0.98	0.58	1426	0.49	0.79	5.94	0.69

Enzyme peptide	GLY	periphery	c	3.75	2.22	875	0.68	3.22	9.88	1.12
Enzyme peptide	ALA	periphery	a	4.60	2.44	676	0.68	5.31	8.93	0.93
Enzyme peptide	SER	periphery	d	3.44	2.14	733	0.58	2.85	8.39	0.92
Enzyme peptide	PRO	periphery	b	4.10	1.94	621	0.58	4.21	7.97	3.53
Enzyme peptide	VAL	periphery	de	3.28	2.22	568	0.65	2.58	8.01	0.92
Enzyme peptide	THR	periphery	ef	3.05	2.09	569	0.59	2.34	8.80	0.78
Enzyme peptide	CYS	periphery	def	3.20	2.22	175	0.65	2.55	7.77	0.73
Enzyme peptide	ILE	periphery	fgh	2.79	2.03	482	0.61	1.78	8.00	0.89
Enzyme peptide	LEU	periphery	fg	2.97	2.22	678	0.49	1.90	8.72	0.79
Enzyme peptide	ASN	periphery	j	2.33	1.70	514	0.55	1.68	7.85	0.86
Enzyme peptide	ASP	periphery	h	2.69	2.01	898	0.56	1.87	7.74	0.87
Enzyme peptide	GLN	periphery	hij	2.56	1.88	524	0.53	1.72	7.79	0.76
Enzyme peptide	LYS	periphery	h	2.74	2.06	682	0.55	1.83	8.67	0.87
Enzyme peptide	GLU	periphery	fgh	2.86	1.97	718	0.55	2.27	7.63	0.78
Enzyme peptide	MET	periphery	ij	2.48	1.91	272	0.61	1.60	7.76	0.65
Enzyme peptide	HIS	periphery	jklmn	2.00	1.78	453	0.50	1.10	7.08	0.65
Enzyme peptide	PHE	periphery	hi	2.63	2.07	580	0.53	1.45	9.78	0.84
Enzyme peptide	ARG	periphery	hij	2.56	2.02	948	0.52	1.66	9.05	0.97
Enzyme peptide	TYR	periphery	jklm	2.02	1.83	683	0.49	1.13	7.78	0.66
Enzyme peptide	TRP	periphery	jklmn	2.00	1.74	277	0.50	1.11	7.17	0.60
Protein peptide	GLY	periphery	b	4.04	2.04	1484	0.75	4.04	9.28	2.30
Protein peptide	ALA	periphery	b	4.06	2.00	1439	0.55	4.01	8.45	3.94
Protein peptide	SER	periphery	d	3.39	1.95	1541	0.63	3.01	7.93	1.58
Protein peptide	PRO	periphery	bc	3.95	1.75	1291	0.61	4.26	8.35	4.58
Protein peptide	VAL	periphery	f	3.02	1.95	1387	0.59	2.40	7.88	0.90
Protein peptide	THR	periphery	fgh	2.86	1.84	1766	0.57	2.18	7.78	0.94
Protein peptide	CYS	periphery	bc	3.96	2.08	140	0.61	4.38	7.72	1.00
Protein peptide	ILE	periphery	gh	2.77	1.86	1038	0.55	1.93	7.24	0.74
Protein peptide	LEU	periphery	ef	3.03	1.88	1603	0.51	2.84	8.11	0.98
Protein peptide	ASN	periphery	hi	2.61	1.73	1339	0.56	1.95	7.65	0.70
Protein peptide	ASP	periphery	fgh	2.87	1.80	1428	0.55	2.43	7.86	1.22
Protein peptide	GLN	periphery	ij	2.41	1.63	1138	0.56	1.91	7.28	0.87
Protein peptide	LYS	periphery	hi	2.60	1.75	1902	0.52	2.08	7.41	0.78
Protein peptide	GLU	periphery	h	2.69	1.85	1776	0.57	1.99	8.24	0.84

Protein peptide	MET	periphery	gh	2.76	1.99	485	0.53	1.89	7.48	0.97
Protein peptide	HIS	periphery	j	2.36	1.63	680	0.57	1.82	7.49	0.88
Protein peptide	PHE	periphery	gh	2.76	1.92	1006	0.53	2.28	8.06	0.61
Protein peptide	ARG	periphery	h	2.68	1.94	2130	0.51	1.89	7.70	0.73
Protein peptide	TYR	periphery	mnop	1.86	1.58	2047	0.49	1.16	7.24	0.83
Protein peptide	TRP	periphery	opq	1.65	1.51	978	0.49	0.79	6.52	0.64
Non-identical	GLY	periphery	cd	3.58	1.45	48538	0.64	3.55	9.47	3.43
Non-identical	ALA	periphery	fg	2.92	1.35	46185	0.55	2.72	8.84	2.38
Non-identical	SER	periphery	h	2.75	1.36	58114	0.53	2.52	9.25	1.53
Non-identical	PRO	periphery	gh	2.79	1.32	46430	0.50	2.62	8.68	1.51
Non-identical	VAL	periphery	j	2.31	1.27	42914	0.53	1.96	8.77	1.06
Non-identical	THR	periphery	j	2.38	1.26	51515	0.45	2.10	8.25	1.16
Non-identical	CYS	periphery	ij	2.47	1.39	7102	0.57	2.12	9.05	1.21
Non-identical	ILE	periphery	jk	2.09	1.21	38951	0.45	1.74	8.48	0.98
Non-identical	LEU	periphery	jkl	2.04	1.16	70542	0.42	1.68	9.41	1.09
Non-identical	ASN	periphery	j	2.25	1.23	46999	0.42	1.92	8.31	1.01
Non-identical	ASP	periphery	j	2.31	1.25	59342	0.51	2.02	8.61	0.94
Non-identical	GLN	periphery	jkl	2.04	1.17	48949	0.46	1.73	8.71	0.77
Non-identical	LYS	periphery	jk	2.10	1.24	69125	0.41	1.75	8.31	0.76
Non-identical	GLU	periphery	j	2.16	1.20	75013	0.45	1.84	8.49	1.19
Non-identical	MET	periphery	jk	2.11	1.27	20728	0.48	1.72	8.34	1.09
Non-identical	HIS	periphery	jklmn	1.97	1.18	24802	0.46	1.59	7.94	1.12
Non-identical	PHE	periphery	klmnop	1.96	1.19	35589	0.46	1.56	8.58	0.96
Non-identical	ARG	periphery	lmnop	1.91	1.15	79750	0.42	1.55	8.49	0.98
Non-identical	TYR	periphery	nop	1.74	1.12	45797	0.39	1.33	8.18	0.79
Non-identical	TRP	periphery	nop	1.75	1.09	15175	0.49	1.40	7.59	0.70
Identical symmetric	GLY	periphery	d	3.35	1.39	39409	0.66	3.33	8.66	4.29
Identical symmetric	ALA	periphery	h	2.70	1.23	38679	0.58	2.51	9.05	1.83
Identical symmetric	SER	periphery	ij	2.50	1.25	40503	0.55	2.26	8.21	1.40
Identical symmetric	PRO	periphery	ij	2.51	1.20	36922	0.47	2.33	8.68	1.42
Identical symmetric	VAL	periphery	jkl	2.08	1.11	32036	0.51	1.79	7.55	0.99

Identical symmetric	THR	periphery	j	2.15	1.14	37075	0.52	1.86	8.76	1.15
Identical symmetric	CYS	periphery	j	2.17	1.16	4857	0.59	1.86	7.24	1.18
Identical symmetric	ILE	periphery	mnop	1.85	1.03	25907	0.47	1.54	7.02	0.99
Identical symmetric	LEU	periphery	nop	1.81	0.98	51008	0.46	1.52	8.73	0.99
Identical symmetric	ASN	periphery	jkl	2.08	1.11	34752	0.52	1.78	7.34	1.02
Identical symmetric	ASP	periphery	jk	2.12	1.12	47534	0.46	1.83	8.58	1.12
Identical symmetric	GLN	periphery	klmnop	1.91	1.04	35606	0.40	1.63	8.62	0.82
Identical symmetric	LYS	periphery	klmnop	1.94	1.09	51493	0.49	1.64	8.21	0.90
Identical symmetric	GLU	periphery	jklmn	1.98	1.06	61640	0.36	1.71	8.74	0.86
Identical symmetric	MET	periphery	jkl	2.04	1.27	14669	0.43	1.65	7.76	0.91
Identical symmetric	HIS	periphery	mnop	1.86	1.09	23441	0.49	1.52	7.70	0.85
Identical symmetric	PHE	periphery	nop	1.74	1.02	25200	0.48	1.42	7.49	0.82
Identical symmetric	ARG	periphery	opq	1.69	0.97	65706	0.47	1.38	7.61	0.76
Identical symmetric	TYR	periphery	opq	1.63	0.99	29253	0.39	1.29	8.09	0.78
Identical symmetric	TRP	periphery	pq	1.51	0.95	9731	0.47	1.16	6.84	0.78
Identical non-symmetric	GLY	periphery	cd	3.53	1.38	17337	0.69	3.53	8.36	4.58
Identical non-symmetric	ALA	periphery	fgh	2.84	1.28	17197	0.61	2.66	9.05	2.23
Identical non-symmetric	SER	periphery	h	2.66	1.27	18760	0.57	2.46	8.13	1.55
Identical non-symmetric	PRO	periphery	hi	2.60	1.22	14483	0.54	2.44	8.56	1.56
Identical non-symmetric	VAL	periphery	j	2.22	1.18	15209	0.46	1.90	8.08	0.94
Identical non-symmetric	THR	periphery	j	2.31	1.20	18062	0.52	2.06	9.29	1.05
Identical non-symmetric	CYS	periphery	j	2.22	1.16	1765	0.57	1.96	7.47	1.08
Identical non-symmetric	ILE	periphery	klmnop	1.95	1.07	13824	0.52	1.65	7.40	1.30
Identical non-symmetric	LEU	periphery	lmnop	1.89	1.07	23299	0.49	1.56	8.05	0.88

Identical non-symmetric	ASN	periphery	j	2.23	1.17	17005	0.51	1.94	8.11	1.47
Identical non-symmetric	ASP	periphery	j	2.21	1.17	21003	0.46	1.92	8.73	1.17
Identical non-symmetric	GLN	periphery	jklm	2.00	1.08	16999	0.46	1.71	7.88	1.02
Identical non-symmetric	LYS	periphery	jklmn	1.97	1.10	22692	0.46	1.67	8.71	0.82
Identical non-symmetric	GLU	periphery	jkl	2.06	1.12	25800	0.46	1.77	8.67	0.95
Identical non-symmetric	MET	periphery	jklmn	1.99	1.17	6794	0.50	1.63	8.30	0.89
Identical non-symmetric	HIS	periphery	klmnop	1.94	1.14	7541	0.45	1.59	9.06	0.79
Identical non-symmetric	PHE	periphery	nop	1.80	1.05	10769	0.43	1.45	8.57	0.75
Identical non-symmetric	ARG	periphery	nop	1.75	1.02	26357	0.47	1.45	8.59	0.81
Identical non-symmetric	TYR	periphery	opq	1.68	1.04	11846	0.37	1.32	8.20	0.74
Identical non-symmetric	TRP	periphery	pq	1.57	0.96	3817	0.39	1.24	6.69	0.70

Table S56 - Summary statistics and ANOVA analysis of hotspots per 100 Å² BSA by interface segmentation. ANOVA p-value < 0.05.

Segmentation	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Single	a	0.87	0.36	10694	0	0.87	3.74	0
Multi	b	0.80	0.32	44415	0	0.81	3.33	0

Table S57 - Summary statistics and ANOVA analysis of hotspots per 100 Å² BSA by interface type. ANOVA p-value < 0.05.

Interface type	ANOVA Group	Mean	SD	N	Min	Med	Max	Mode
Enzyme-peptide	a	0.97	0.35	822	0.17	0.95	2.25	0.17
Protein-peptide	b	1.04	0.33	1702	0.13	1.07	2.34	0.13
Non-identical	c	0.77	0.33	28165	0	0.77	3.49	0.00
Identical-symmetric	d	0.90	0.30	15920	0	0.92	3.74	0.00
Identical-nonsymmetric	e	0.73	0.32	8580	0	0.73	3.46	0.00

Table S58 - Residue exposure based on distance from interacting chain (s) and change in residue side-chain relative solvent accessibility (RSA) on complexation. A residue is considered to be at the interface if at least one of its atoms is within 5 Å of any of the binding partner's protein atoms.

Residue % RSA (Single chain)	Residue % RSA (Whole complex)	Is interface?	Exposure category

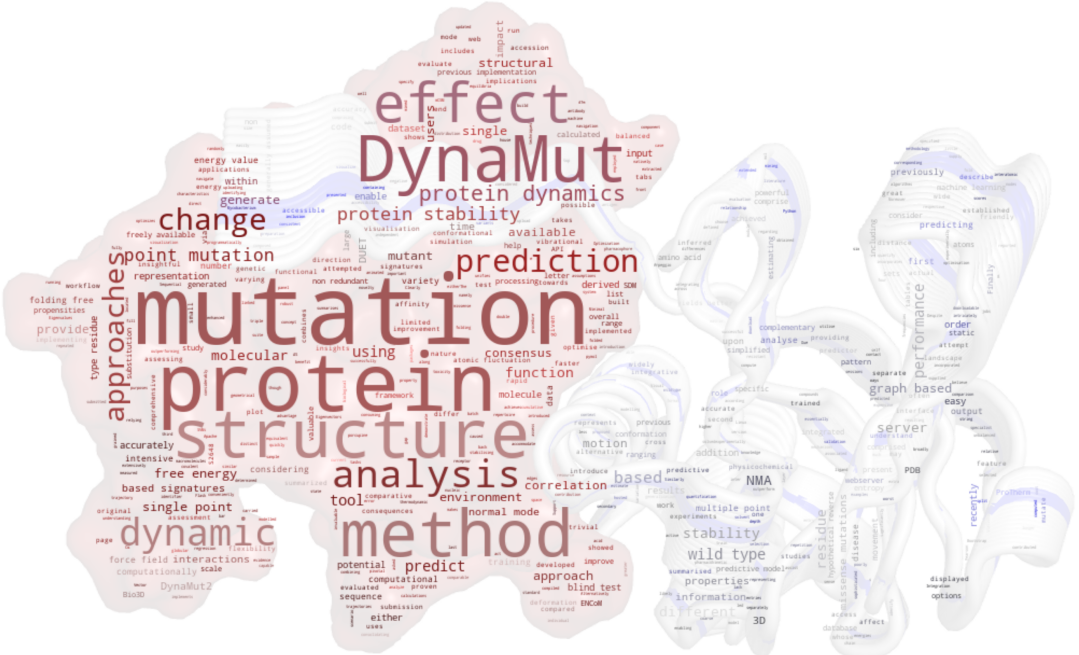
≤ 7	Any	Any	Protein Core
> 7	Any	False	Surface exposed
> 7	≤ 7	True	Interface Core
> 7	> 7	True	Interface Periphery

SI References

1. Lesk, A., *Introduction to Bioinformatics*. 2019: Oxford university press.

Chapter 3

Predicting the Effects of Mutations on Protein Conformation, Flexibility and Stability



W350–W355 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 30 April 2018*
doi: 10.1093/nar/gky300

DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability

Carlos H.M. Rodrigues[1], Douglas E.V. Pires[2,*] and David B. Ascher[1,2,3,*]

[1]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Australia, [2]Instituto René Rachou, Fundação Oswaldo Cruz, Brazil and [3]Department of Biochemistry, University of Cambridge, UK

Received January 31, 2018; Revised April 03, 2018; Editorial Decision April 08, 2018; Accepted April 16, 2018

ABSTRACT

Proteins are highly dynamic molecules, whose function is intrinsically linked to their molecular motions. Despite the pivotal role of protein dynamics, their computational simulation cost has led to most structure-based approaches for assessing the impact of mutations on protein structure and function relying upon static structures. Here we present DynaMut, a web server implementing two distinct, well established normal mode approaches, which can be used to analyze and visualize protein dynamics by sampling conformations and assess the impact of mutations on protein dynamics and stability resulting from vibrational entropy changes. DynaMut integrates our graph-based signatures along with normal mode dynamics to generate a consensus prediction of the impact of a mutation on protein stability. We demonstrate our approach outperforms alternative approaches to predict the effects of mutations on protein stability and flexibility (*P*-value < 0.001), achieving a correlation of up to 0.70 on blind tests. DynaMut also provides a comprehensive suite for protein motion and flexibility analysis and visualization via a freely available, user friendly web server at http://biosig.unimelb.edu.au/dynamut/.

INTRODUCTION

Proteins are dynamic macromolecules, whose function is intricately linked to their biological motions (1,2). We have shown previously that drug resistant and genetic disease mutations can both act through changes in protein conformational equilibria and dynamics (3–7). In order to fully understand the molecular consequences of a mutation it is, therefore, important to consider changes in protein dynamics. Despite their pivotal role, the computational cost of dynamics simulation has led to most structure-based approaches for assessing mutations effects on protein structure and function relying upon static structures.

Normal Mode Analysis (NMA) is a computational approach that approximates the dynamics of a system around a conformation through harmonic motion. This has been used to generate possible movements and therefore provide valuable insights into protein motions, and their accessible conformational repertoires. Previous studies have shown that NMA can be a powerful tool to analyze protein structure–function relationship (8) and to predict the effects of single-point mutations on protein stability (9). Many NMA methods have been proposed (10–14) to address the lack of easy to use interfaces that limited their use to those with specialist knowledge. However, these are limited to the analysis of protein structures and do not provide approaches to evaluate the effect of mutations within their pipelines.

To fill this gap, we introduce DynaMut, a web server that introduces the dynamics component to mutation analysis. This is achieved by implementing and integrating well established normal mode approaches with our graph-based signatures in a consensus predictor for protein stability changes upon mutation, which we show optimizes overall prediction performance.

DynaMut implements NMA through two different approaches, Bio3D (8) and ENCoM (9), providing rapid and simplified access to powerful and insightful analysis of protein motions. In addition, DynaMut also enables rapid analysis of the impact of mutations on a protein's dynamics and stability resulting from vibrational entropy changes. Integration of these two different approaches with other well-established methods and characteristics of the wild-type residue environment into a consensus prediction enables DynaMut to provide an accurate assessment of the impact of a mutation on protein stability, and provide a comprehensive suite for protein motion and flexibility analysis and visualization via an easy-to-use web interface (http://biosig.unimelb.edu.au/dynamut/).

*To whom correspondence should be addressed. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au
Correspondence may also be addressed to Douglas E. V. Pires. Email: douglas.pires@minas.fiocruz.br

MATERIALS AND METHODS

Data sets

In this work, we used the previously established S2648 dataset (15–18), derived from the ProTherm database (19). This dataset is comprised of 2648 different point-mutations across 131 globular proteins with experimentally determined structures whose impact on protein stability has been experimentally measured (602 stabilizing and 2046 destabilizing). The DynaMut training set comprises 2297 mutations randomly selected from the original dataset. A blind test set composed of 351 non-redundant mutations derived from the S2648 set was also compiled. This blind test set has been widely used in the literature (15–18), enabling direct comparative performance of methods that quantify the impact of mutations on the folding free energy.

Previous studies have reported performance comparisons of difference methods on predicting changes in folding free energy ($\Delta\Delta G$) using these datasets (20–22). Given the unbalanced nature of the original dataset, here we have considered the hypothetical reverse mutations (22) in order to build a more robust, balanced and self-consistent predictive method. The change in folding free energy is a thermodynamic state function, and it has been proposed that the change in folding free energy of a mutation from a wild-type protein to its mutant ($\Delta\Delta G_{WT\rightarrow MT}$) should be equivalent to the negative change in folding free energy of the hypothetical reverse mutation—from the mutant to the wild-type protein ($-\Delta\Delta G_{MT\rightarrow WT}$) (16,22–24). Including the hypothetical reverse mutations, our predictive model was trained using 4594 mutations and our blind test was comprised of 702 single-point mutations.

Normal mode analysis

NMA allows the study of harmonic motions in a system, providing insights into its dynamics and accessible conformations. It has been widely used for studies of protein dynamics as an alternative to more computationally intensive molecular dynamics approaches (25–28). While molecular dynamics approaches provide motion trajectories for a given molecule over time, conformational fluctuations can be evaluated by NMA via superposition of normal modes (Eigenvectors) and their associated frequencies (Eigenvalues) (29). NMA can also use simplified representations of the protein structure, such as modeling the amino acids using their Cα atoms, reducing computational cost. NMA has been successfully applied to the study of the effects of mutations on protein dynamics, with ENCoM (9) including the nature of the amino acids in the protein as an extra layer of information to compute the effects of single-point mutations on the vibrational entropy (ΔS) and protein stability.

Other structure-based approaches

Structure-based approaches to predict the impact of mutations on stability utilize protein structural information from the 3D space of a natively folded protein. Even though these structure-based methods are essentially based on the same structural data, they are built using broadly different, sophisticated, approaches, such as statistical potential function energy calculations, used in SDM (16) and structural pattern mining approaches such as mCSM-Stability (18). The consensus method DUET highlighted that these approaches were complimentary, and that their integration provided more accurate and reliable predictions (17). This has been used to provide invaluable insights into disease and drug resistance mutations, and help guide protein engineering efforts (30–39).

DynaMut—consensus predictions

Within DynaMut we have implemented a consensus estimate of changes upon mutation on protein folding free energy, which combines the effects of mutations on protein stability and dynamics calculated by Bio3D, ENCoM and DUET to generate an optimized and more robust predictor. Moreover, DynaMut includes a set of complementary information regarding the environment characteristics of the wild-type residue (*e.g.*, relative solvent accessibility, residue depth and secondary structure) and graph-based signatures representing the wild-type structure. The graph-based signatures concept, used in the development of mCSM-Stability and to generate the consensus DUET predictions, has been widely applied to the study of protein structure, including protein–ligand interactions (40), and how mutations alter protein interactions with other molecules (23,24,41–43). These were supplied as evidence for training the consensus predictor using Random Forest (44). Figure 1 shows the workflow used to train the consensus predictions. The DynaMut consensus prediction was trained under 10-fold cross validation, and validated using the non-redundant blind test set (Supplementary Materials). The machine learning algorithm, evaluation procedures, performance metrics and details on the methods used on the consensus prediction are described in Supplementary Materials.

WEB SERVER

We have implemented DynaMut as a user-friendly, freely available web server (http://biosig.unimelb.edu.au/dynamut/). The server front end was built using Bootstrap framework version 3.3.7, while the back-end was built in Python via the Flask framework (Version 0.12.2). It is hosted on a Linux server running Apache.

Input

DynaMut can be used in two different ways, to either (1) analyze protein dynamics or (2) to analyze the effect of point mutations on protein dynamics and stability. For protein dynamics analysis (Supplementary Figure S1), the server requires the user to input a protein structure by either uploading a file in PDB format or by providing the four-letter accession code for any entry on the PDB database. In addition, users have the option to choose a specific force field, which is used to describe the molecular interactions within the structure for normal mode analysis. The force field options available are summarized in Supplementary Table S1 of Supplementary Materials.

Alternatively, for assessing the effects of mutations on protein dynamics and stability, two different input options

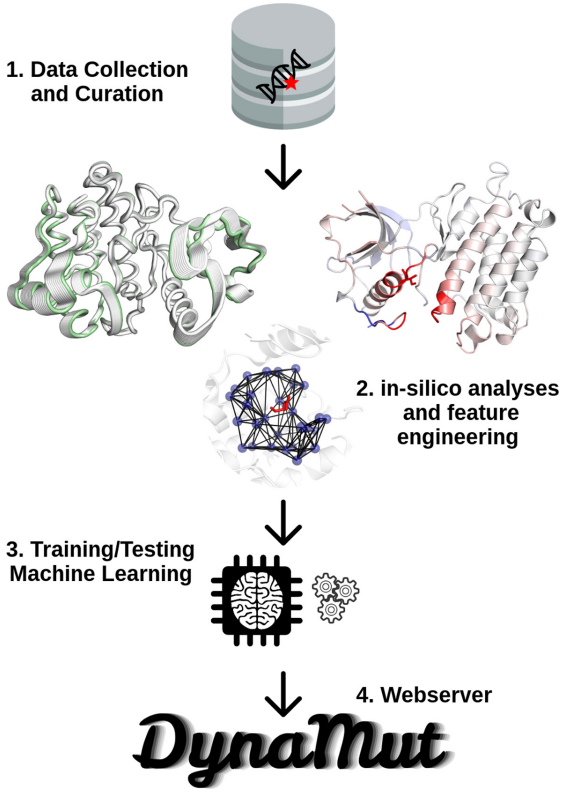


Figure 1. Methodology workflow. The DynaMut methodology can be divided into four steps. In step 1, data was collected from the previously established S2648 subset of mutations with experimental evidence from ProTherm. In step 2, DynaMut combines the effects of mutations on protein stability and dynamics calculated by Bio3D, ENCoM and DUET. In addition, DynaMut also includes a set of complementary information regarding the environment characteristics of the wild-type residue (e.g. relative solvent accessibility, residue depth and secondary structure) and the graph-based signatures generated by mCSM. All these features are used as evidence for training supervised learning algorithms in step 3. After evaluating the performance of the predictive model, the consensus prediction was integrated into the DynaMut web server.

are available (Supplementary Figure S2). The 'Single mutation' option requires the user to provide a PDB file or PDB accession code, the point mutation specified as a string containing the wild-type residue one-letter code, its corresponding residue number and the mutant residue one-letter code. The 'Mutation list' option allows users to upload a list of mutations in a file for batch processing. For both input options the user is also asked to specify the chain identifier in which the wild-type residue is located.

In order to assist users to submit their jobs for analysis and predictions, sample submission entries are available in both submission pages and a help page is available via the top navigation bar.

Output

For the analysis of protein dynamics, the results are displayed in four tabs. In the first tab (Supplementary Figure S3), porcupine plots show the trajectory of movement according to the first non-trivial mode of the molecule. The second tab (Supplementary Figure S4) allows users to visualize the non-trivial modes generated, including an animated plot that describes the motion of the molecule. Visual representations of deformation energy and atomic fluctuation are displayed on the third tab (Supplementary Figure S5). Finally, the last tab shows the cross-correlation between residue movements as both a correlation matrix and the 3D structure of the submitted protein (Supplementary Figure S6).

The mutational analysis results are also split into tabs to enable users to easily navigate the different analyses available for evaluating the effects of mutations on protein stability and dynamics. For the 'Single mutation' option, the server outputs the predicted change in stability (in kcal/mol), along with the variation in entropy energy between wild-type and mutant structures (in kcal/mol/K) in the first tab (Supplementary Figure S7). For comparison purposes, in a separate panel the changes in stability calculated by structure-based methods are shown (16–18). DynaMut enables visualization of the non-covalent molecular interactions calculated by Arpeggio (45) (Supplementary Table S2, Supplementary Figure S8) and deformation energies and atomic fluctuations of wild-type and mutant residues (Supplementary Table S3, Supplementary Figure S9) in their respective 3D structures. For the 'Mutation list' option, the server output is summarized as a downloadable table, and users have the option to analyze each mutation separately, similar to the analysis of a single mutation (Supplementary Figure S10).

DynaMut also generates and makes available for download pymol sessions for flexibility analysis and for inter-residue interactions for both wild-type and mutant structures to facilitate easy visualisation and figure preparation.

VALIDATION

The performance of DynaMut was compared to well-established methods that also provide measurements of effects of single-point mutations on protein stability. All mutations from the data set described previously were submitted to each tool and the Pearson's Correlation Coefficient and Root Mean Squared Error were used to assess the comparison among all methods. Moreover, outliers were considered based on the absolute difference between predicted and actual values of $\Delta\Delta G$.

Since this definition can vary across the methods and for comparison purposes we defined $\Delta\Delta G \geq 0$ as stabilizing and $\Delta\Delta G < 0$ as destabilizing. In the case that a method does not follow such definition, its results were adapted.

Performance on cross validation

Across the full training set (forward and reverse mutations), DynaMut achieved a Pearson's correlation of $r = 0.67$, and RMSE = 1.31 kcal/mol ($r = 0.79$ and $\sigma = 0.01$ on 90% of the data) under 10-fold cross validation. This correlation was significantly higher than the individual methods used in the consensus prediction (P-value < 0.0001). Supplementary Table S1 on Supplementary Materials summarizes the performance for all the methods during training of DynaMut. Figure 2A shows the regression analysis for performance of DynaMut over the training set.

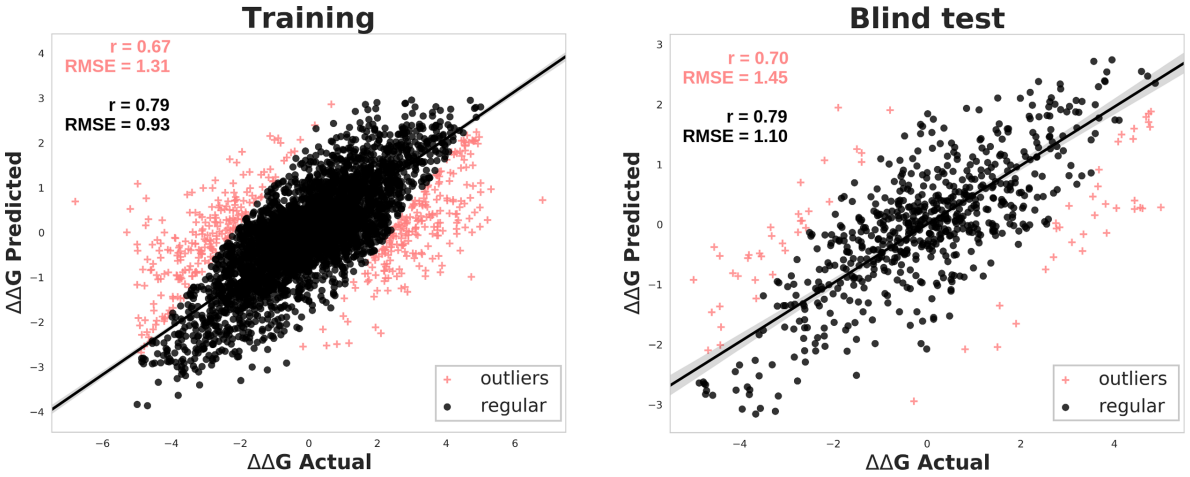


Figure 2. Regression analysis of the performance of DynaMut over training and blind test. Left panel shows the correlation during training and Right panel depicts the correlation between the actual values of $\Delta\Delta G$ and the predictions of DynaMut. Pearson's correlation coefficient (r) and RMSE are shown. Crosses in pink show the 10% outliers. The performance results are shown on the top left of each panel. The results colored in pink are related to the entire dataset and the results colored in black were obtained after removing 10% of the outliers.

Table 1. Performance of DynaMut on Blind test for the 351 mutations with experimental 3D structure (forward), the 351 hypothetical reverse mutations (reverse) and the overall results for all the 702 mutations (Overall). The performance of well-established methods are also shown for comparison purposes

Method	Forward		Reverse		Overall	
	Pearson (r)	RMSE	Pearson (r)	RMSE	Pearson (r)	RMSE
DynaMut	**0.69**	**1.39**	**0.58**	**1.51**	**0.70**	**1.45**
I-Mutant 2 (46)	0.73	1.01	0.21[a]	2.55	0.49[a]	1.97
Maestro (47)	0.20[a]	2.13	0.60	2.12	0.49[a]	2.13
DUET (17)	0.75	1.05	0.27[a]	2.39	0.56[a]	1.85
SDM2 (16)	0.52[a]	1.80	0.42[a]	2.16	0.50[a]	1.99
mCSM (18)	0.76	1.09	0.23[a]	2.50	0.54[a]	1.93
ENCoM (9)	0.44[a]	1.79	−0.50[a]	2.31	0.35[a]	1.79
FoldX (48)	0.35[a]	2.33	−0.29[a]	2.23	−0.55[a]	2.32

[a] P-value < 0.001 compared to DynaMut using z-test.

Blind test

The non-redundant blind test was used to evaluate the generalization of the consensus predictions. Across the complete blind test set of 702 mutations containing both forward and hypothetical reverse mutations, DynaMut obtained a Pearson's correlation coefficient of 0.70 (RMSE = 1.45; Figure 2B). After removing 10% outliers, DynaMut achieves a correlation of up to $r = 0.79$ (RMSE = 1.10; Figure 2B). This was significantly higher (P-value < 0.001) than comparable methods (Table 1).

Looking specifically at those data points with experimental data, the original core 351 non-redundant mutations, DynaMut achieved a Pearson's correlation of $r = 0.69$ (RMSE = 1.39), significantly higher than the performance of either ENCoM, FoldX, SDM or Maestro, but lower than I-Mutant2, DUET and mCSM (P-value < 0.001; Table 1). Considering the hypothetical reverse mutations alone, DynaMut significantly outperformed all other algorithms tested, achieving a Pearson's correlation of 0.58 (RMSE = 1.51; Table 1).

Previous studies have highlighted that many machine learning based structural approaches are unbalanced, and can less accurately identify stabilizing mutations (16). We therefore considered method performance across stabilizing and destabilizing mutations separately (Supplementary Table S2). Considering the destabilizing mutations alone, DynaMut has a comparable correlation coefficient but higher RMSE (1.42) than mCSM (1.02), DUET (1.04) and iMutant2 (1.07), and outperformed the other methods tested. Across the stabilizing mutations, however, DynaMut achieved a correlation of $r = 0.51$ (RMSE = 1.48), significantly higher than all comparative methods ($P < 0.01$; Supplementary Table S3). This highlights that DynaMut provides the most accurate and balanced approach for the prediction of both destabilizing and stabilizing mutations.

CONCLUSION

Here, we present DynaMut, an integrated computational method that provides users with easy access to powerful and insightful analysis of protein motions and their changes upon mutation. By consolidating these insights with our graph-based signatures, DynaMut is able to accurately assess the effects of missense mutations on protein stability. This consensus approach allows for the more accurate and reliable prediction of both stabilizing and destabilizing mutations. DynaMut is a valuable tool for a wide variety of

applications, ranging from protein functional analysis, optimization of stability and understanding the role of mutations in diseases. The method is freely available as a user friendly and easy to use web server at http://biosig.unimelb.edu.au/dynamut/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Australian Government Research Training Program Scholarship [to C.H.M.R]; Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A., D.E.V.P.]; National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; Victorian Life Sciences Computation Initiative (VLSCI), an initiative of the Victorian Government, Australia, on its Facility hosted at the University of Melbourne [UOM0017]; Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P.]; Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.]. Funding for open access charge: MRC.
Conflict of interest statement. None declared.

REFERENCES

1. Karplus,M. and Kuriyan,J. (2005) Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6679–6685.
2. Jubb,H., Blundell,T.L. and Ascher,D.B. (2015) Flexibility and small pockets at protein-protein interfaces: New insights into druggability. *Prog. Biophys. Mol. Biol.*, **119**, 2–9.
3. Albanaz,A.T.S., Rodrigues,C.H.M., Pires,D.E.V. and Ascher,D.B. (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin. Drug Discov.*, **12**, 553–563.
4. Ascher,D.B., Wielens,J., Nero,T.L., Doughty,L., Morton,C.J. and Parker,M.W. (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci. Rep.*, **4**, 4765.
5. Ramdzan,Y.M., Trubetskov,M.M., Ormsby,A.R., Newcombe,E.A., Sui,X., Tobin,M.J., Bongiovanni,M.N., Gras,S.L., Dewson,G., Miller,J.M.L. *et al.* (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. *Cell Rep.*, **19**, 919–927.
6. Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B., Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom. Med.*, **2**, 7.
7. Trezza,A., Bernini,A., Langella,A., Ascher,D.B., Pires,D.E.V., Sodi,A., Passerini,I., Pelo,E., Rizzo,S., Niccolai,N. *et al.* (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest. Ophthalmol. Vis. Sci.*, **58**, 5320–5328.
8. Grant,B.J., Rodrigues,A.P., ElSawy,K.M., McCammon,J.A. and Caves,L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.
9. Frappier,V. and Najmanovich,R.J. (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. *PLoS Comput. Biol.*, **10**, e1003569.
10. Lopez-Blanco,J.R., Aliaga,J.I., Quintana-Orti,E.S. and Chacon,P. (2014) iMODS: internal coordinates normal mode analysis server. *Nucleic Acids Res.*, **42**, W271–W276.
11. Camps,J., Carrillo,O., Emperador,A., Orellana,L., Hospital,A., Rueda,M., Cicin-Sain,D., D'Abramo,M., Gelpi,J.L. and Orozco,M. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics*, **25**, 1709–1710.
12. Suhre,K. and Sanejouand,Y.H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, **32**, W610–W614.
13. Eyal,E., Lum,G. and Bahar,I. (2015) The anisotropic network model web server at 2015 (ANM 2.0). *Bioinformatics*, **31**, 1487–1489.
14. Tiwari,S.P., Fuglebakk,E., Hollup,S.M., Skjaerven,L., Cragnolini,T., Grindhaug,S.H., Tekle,K.M. and Reuter,N. (2014) WEBnm@ v2.0: Web server and services for comparing protein flexibility. *BMC Bioinformatics*, **15**, 427.
15. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
16. Pandurangan,A.P., Ochoa-Montano,B., Ascher,D.B. and Blundell,T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.*, **45**, W229–W235.
17. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–W319.
18. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
19. Kumar,M.D., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
20. Potapov,V., Cohen,M. and Schreiber,G. (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng. Des. Sel.*, **22**, 553–560.
21. Khan,S. and Vihinen,M. (2010) Performance of protein stability predictors. *Hum Mutat*, **31**, 675–684.
22. Thiltgen,G. and Goldstein,R.A. (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One*, **7**, e46084.
23. Pires,D.E. and Ascher,D.B. (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.*, **44**, W469–W473.
24. Pires,D.E. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.
25. Tasumi,M., Takeuchi,H., Ataka,S., Dwivedi,A.M. and Krimm,S. (1982) Normal vibrations of proteins: glucagon. *Biopolymers*, **21**, 711–714.
26. Go,N., Noguti,T. and Nishikawa,T. (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. U.S.A.*, **80**, 3696–3700.
27. Levitt,M., Sander,C. and Stern,P.S. (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, **181**, 423–447.
28. Bahar,I., Lezon,T.R., Bakan,A. and Shrivastava,I.H. (2010) Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.*, **110**, 1463–1497.
29. Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.
30. Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.
31. Usher,J.L., Ascher,D.B., Pires,D.E., Milan,A.M., Blundell,T.L. and Ranganath,L.R. (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: Identification of novel mutations. *JIMD Rep.*, **24**, 3–11.
32. Kano,F.S., Souza-Silva,F.A., Torres,L.M., Lima,B.A., Sousa,T.N., Alves,J.R., Rocha,R.S., Fontes,C.J., Sanchez,B.A., Adams,J.H. *et al.* (2016) The presence, persistence and functional properties of plasmodium vivax duffy binding protein II antibodies are influenced by HLA class II allelic variants. *PLoS Negl. Trop. Dis.*, **10**, e0005177.

33. Nemethova,M., Radvanszky,J., Kadasi,L., Ascher,D.B., Pires,D.E., Blundell,T.L., Porfirio,B., Mannoni,A., Santucci,A., Milucci,L. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *E.ur J. Hum. Genet.*, **24**, 66–72.
34. Phelan,J., Coll,F., McNerney,R., Ascher,D.B., Pires,D.E., Furnham,N., Coeck,N., Hill-Cawthorne,G.A., Nair,M.B., Mallard,K. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.*, **14**, 31.
35. White,R.R., Ponsford,A.H., Weekes,M.P., Rodrigues,R.B., Ascher,D.B., Mol,M., Selkirk,M.E., Gygi,S.P., Sanderson,C.M. and Artavanis-Tsakonas,K. (2016) Ubiquitin-Dependent modification of skeletal muscle by the parasitic nematode, trichinella spiralis. *PLoS Pathog.*, **12**, e1005977.
36. Casey,R.T., Ascher,D.B., Rattenberry,E., Izatt,L., Andrews,K.A., Simpson,H.L., Challis,B., Park,S.M., Bulusu,V.R., Lalloo,F. *et al.* (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol. Genet. Genomic Med.*, **5**, 237–250.
37. Pandurangan,A.P., Ascher,D.B., Thomas,S.E. and Blundell,T.L. (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.*, **45**, 303–311.
38. Park,Y., Pacitto,A., Bayliss,T., Cleghorn,L.A., Wang,Z., Hartman,T., Arora,K., Ioerger,T.R., Sacchettini,J., Rizzi,M. *et al.* (2017) Essential but not Vulnerable: Indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against mycobacterium tuberculosis. *ACS Infect. Dis.*, **3**, 18–33.
39. Singh,V., Donini,S., Pacitto,A., Sala,C., Hartkoorn,R.C., Dhar,N., Keri,G., Ascher,D.B., Mondesert,G., Vocat,A. *et al.* (2017) The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. *ACS Infect. Dis.*, **3**, 5–17.
40. Pires,D.E. and Ascher,D.B. (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res.*, **44**, W557–W561.
41. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res.*, **43**, D387–D391.
42. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.*, **6**, 29575.
43. Pires,D.E., Chen,J., Blundell,T.L. and Ascher,D.B. (2016) In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.*, **6**, 19848.
44. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
45. Jubb,H.C., Higueruelo,A.P., Ochoa-Montano,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.
46. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
47. Laimer,J., Hofer,H., Fritz,M., Wegenkittl,S. and Lackner,P. (2015) MAESTRO–multi agent stability prediction upon point mutations. *BMC Bioinformatics*, **16**, 116.
48. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.

SUPPLEMENTARY MATERIAL

DynaMut: analysis and prediction of protein stability changes upon mutation using Normal Mode Analysis

Carlos H.M. Rodrigues[1], Douglas E.V. Pires[3,*], David B. Ascher[1,2,3,*]

[1]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne;
[2]Department of Biochemistry, University of Cambridge;
[3]Instituto René Rachou, Fundação Oswaldo Cruz

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au or da382@cam.ac.uk. Correspondence may also be addressed to D.E.V.P. douglas.pires@minas.fiocruz.br.

CONSENSUS PREDICTION COMPONENTS

Bio3D

Bio3D is a R package that contains utilities that helps one to process, organize and explore protein structure and sequence data. Among other features available, Bio3D provides the ability to read, write and process biomolecular structure, sequence and dynamics trajectory data; perform ensemble normal mode analysis on large structure sets to explore evolutionary dynamics and structure sets to explore evolutionary dynamics and structure dependent protein flexibility; and also various utility functions are provided to enable the statistical and graphical power of the R environment when working with biological sequence and structural data (1). The package source code is freely available at https://bitbucket.org/Grantlab/bio3d/.

ENCoM

ENCoM is an Elastic Network Contact Model that employs a potential energy function and includes a pairwise atom-type non-bonded interaction term to add an extra layer of information regarding the effect of the specific nature of amino acids on dynamics within the context of NMA (2). ENCoM tries to approximate ΔΔG through the calculations of the vibrational entropy (ΔS) (3) of wild-type and mutant structures. The ΔS between two conformations (A, B) in terms of their respective sets of eigenvalues is given by:

$$\Delta S_{Vib,A\rightarrow B} = \ln\left(\frac{\prod_{n=7}^{3N}\lambda_{n,A}}{\prod_{n=7}^{3N}\lambda_{n,B}}\right)$$

where:

- $\boldsymbol{\lambda_{n,i}}$ represents the *n*th normal mode (the first 6 modes correspond to rotational and translational degrees of freedom and because of that they are not considered on the calculations.

The source code for ENCoM is publicly available at https://github.com/NRGlab/ENCoM.

DUET

DUET is an integrated approach for predicting the effects of mutations on protein stability that takes advantage of two distinct techniques, SDM (4) and mCSM (5), by combining them in a consensus prediction (6). DUET unifies the results of the separate methods in an optimised predictor using Support Vector Machines (SVM) trained with Sequential Minimal Optimization (7). DUET predictions were more accurate than either method on their own. DUET is freely available as a web server at http://biosig.unimelb.edu.au/duet.

EVALUATION METRICS

A set of well-established and widely used performance metrics for evaluation regression models were used to evaluate DynaMut on both 10-fold cross validation and on blind tests. These metrics include Pearson's Coefficient of Correlation (*r)* and Root Mean Squared Error (RMSE).

Pearson's Coefficient of Correlation

The Pearson correlation coefficient, also known as the product moment correlation coefficient, is a measure of the linear correlation between two variables X and Y. The coefficient is measure on a scale with no units and can take values from -1, total negative linear correlation, to +1, total positive correlation. Values closer from 0 indicate that there is no linear correlation between X and Y. The mathematical definition of the Pearson's Correlation Coefficient is given by the covariance of the two variables divided by the product of their standard deviations as described by the formula below (8).

$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ where:

- $\boldsymbol{cov(X,Y)}$ is the covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$;
- $\boldsymbol{\sigma_X}$ is the standard deviation of the variable $\boldsymbol{X}$;
- $\boldsymbol{\sigma_Y}$ is the standard deviation of the variable $\boldsymbol{Y}$;

Root Mean Squared Error

Root Mean Squared Error (RMSE) is the standard deviation of the predictions errors. This measure indicates how concentrated the predicted data points are from the line of best fit which represents the ideal perfect correlation between the actual observed values (Y) and the predicted values ($\hat{Y}$) (9). RMSE is described by the formula below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

where:

- $\boldsymbol{n}$ is the total number of instances;
- $(\boldsymbol{Y_i - \hat{Y}_i})^{\boldsymbol{2}}$ represents the squared errors between actual observed values and the predictions;

MACHINE LEARNING

The Machine learning task used on this work was implemented on the Weka Tool Kit (10).

Random Forest

The Random Forest algorithm uses a set of decision tree predictors in a way that each tree relies on the values of a random vector sampled independently and with the same distribution for all trees in the set. The generalization error for forests converge to a limit as the number of trees in the set becomes large (11).

This is a fast and easy to implement algorithm that produces highly accurate predictions and can handle a large number of input variables with low overfitting rates. Since all the trees are built from scratch without any previous information on the other trees (reason why the algorithm is called Random Forest) in the forest and also the final prediction is the average of all the predictions for each tree.

TABLES

Table S1 – Force Fields options for Normal Mode Analysis in DynaMut.

Name	Description
C-alpha (12)	Force field derived from fitting to the Amber94 all-atom potential.
ANM (13)	Anisotropic Network Model uses a simplified spring force constant based on the pair-wise distance.
pfANM (14)	parameter-free Anisotropic Network Model is variant from the ANM force field with interactions that fall off with the square of the distance.
REACH (15)	Realistic Extension Algorithm via Covariance Hessian is parameterized based on variance-covariance matrices obtained from MD simulations.
sdENM (16)	This force field employs residue specific spring force constants and it has been parameterized through a statistical analysis of 1500 NMR ensembles.

Table S2 – Performance evaluation of DynaMut on training and comparison with other methods.

Methods	Pearson (r)	RMSE
DynaMut	**0.67**	**1.31**
DUET (6)	0.41*	1.79
SDM2 (4)	0.42*	1.93
mCSM (5)	0.40*	1.83
ENCoM (2)	0.05*	5.13
FoldX (17)	-0.05*	4.37

** p-value < 0.001 compared to DynaMut using z-test.*

Table S3 – Performance evaluation of DynaMut on identifying stabilizing mutations and comparison with other methods.

Method	Stabilising		Destabilizing	
	Person (*r*)	RMSE	Pearson (*r*)	RMSE
DynaMut	**0.51**	**1.48**	**0.61**	**1.42**
I-Mutant (18)	0.07*	2.57	0.57	1.07
Maestro (19)	0.43	2.22	0.45*	2.13
DUET (6)	0.13*	2.4	0.64	1.04
SDM2 (4)	0.26*	2.15	0.4*	1.83
mCSM (5)	0.12*	2.53	0.63	1.02
ENCoM (2)	0.37*	1.84	0.03*	4.36
FoldX (17)	-0.37*	2.34	-0.03*	5.21

** p-value < 0.01 compared to DynaMut using z-test.*

FIGURES

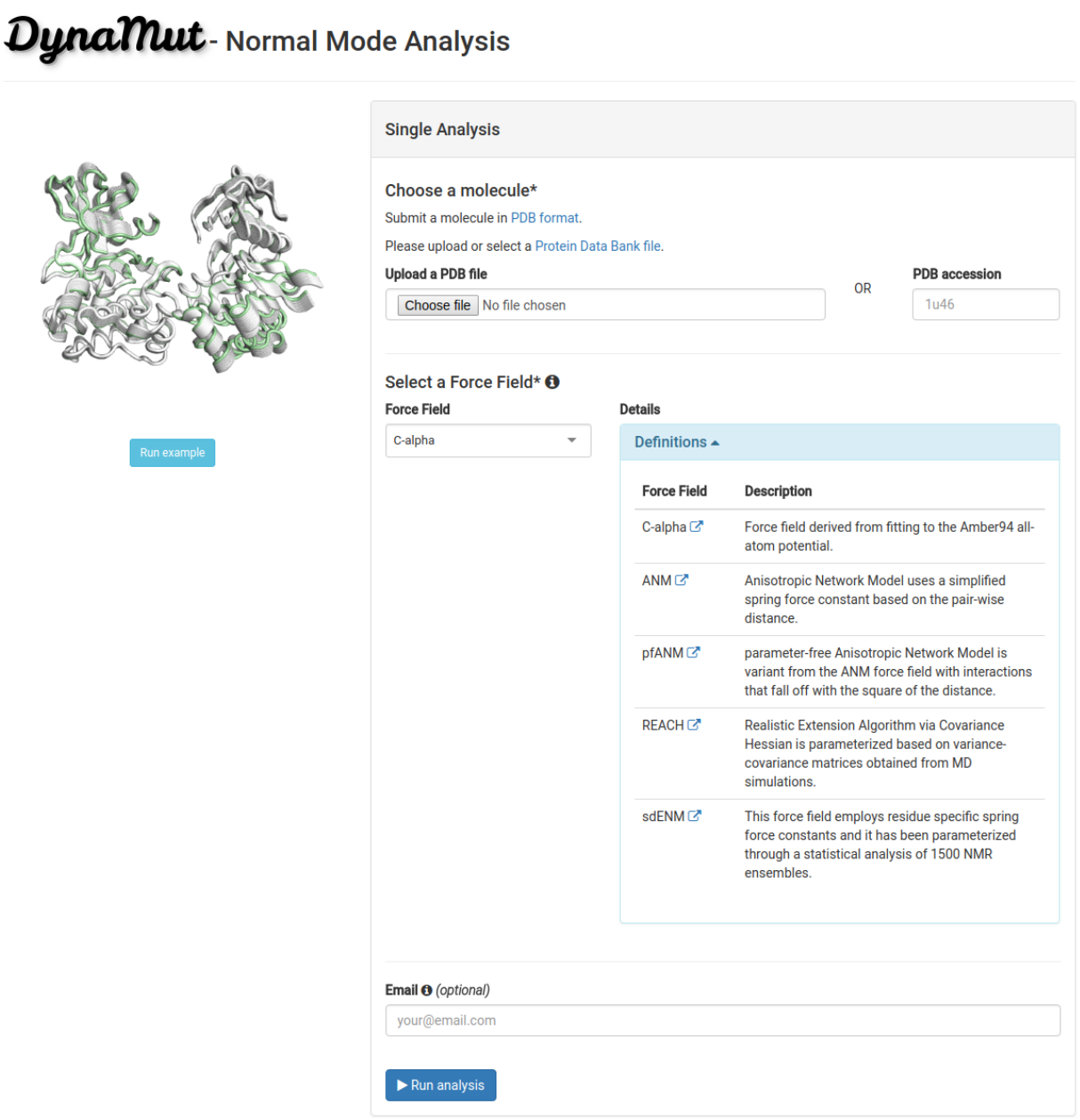


Figure S1 - DynaMut normal mode analysis input page. For the protein dynamics analysis, the server requires the user to input a protein structure by either uploading a file in PDB format or by providing the 4-letter accession code for any entry on the PDB database. In addition, users, are required to specify a force field that will describe the interactions between the atoms of the structure for the normal mode analysis.

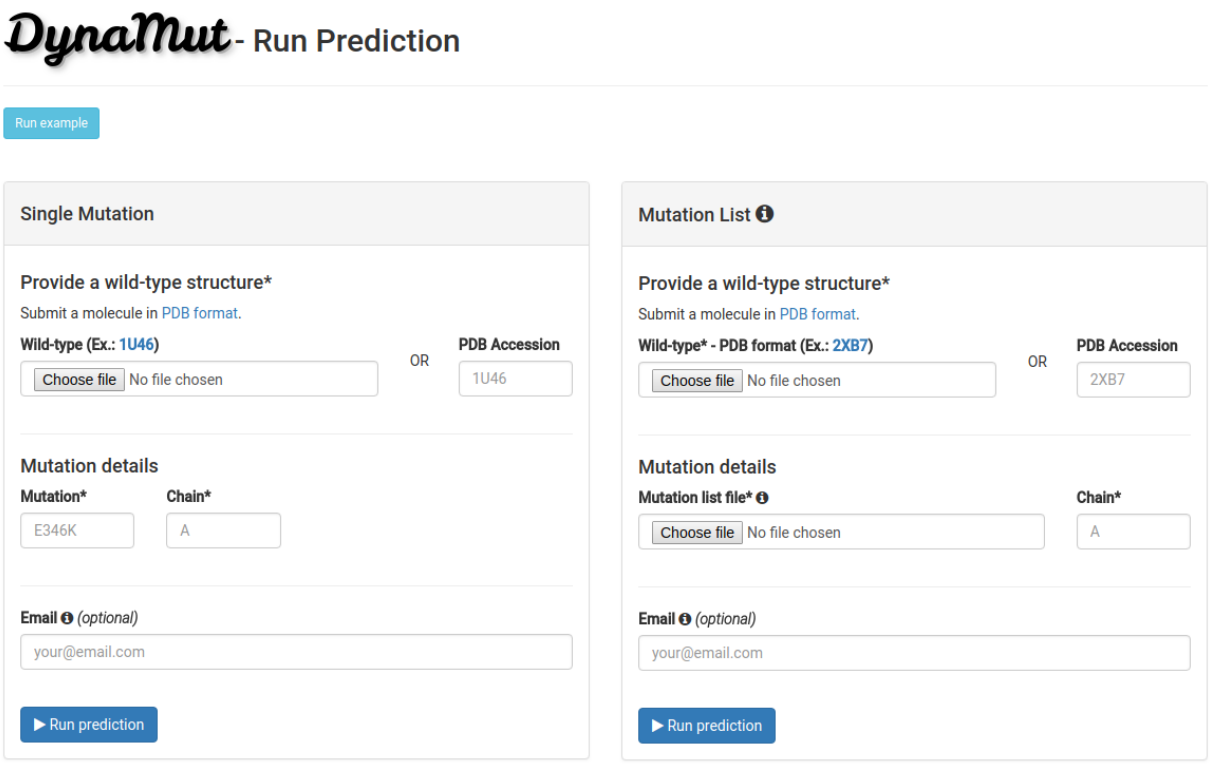


Figure S2 - DynaMut prediction input page. For assessing effects of mutations on protein dynamics and stability two different input options are available. The "Single mutation" option requires the user to provide a PDB file or PDB accession code, the point mutation specified as a string containing the wild-type residue one-letter code, its corresponding residue number and the mutant residue one-letter code. The "Mutation list" option allows users to upload a list of mutations in a file for batch processing. For both input options the user also is asked to specify the chain identifier in which the wild-type residue is located.

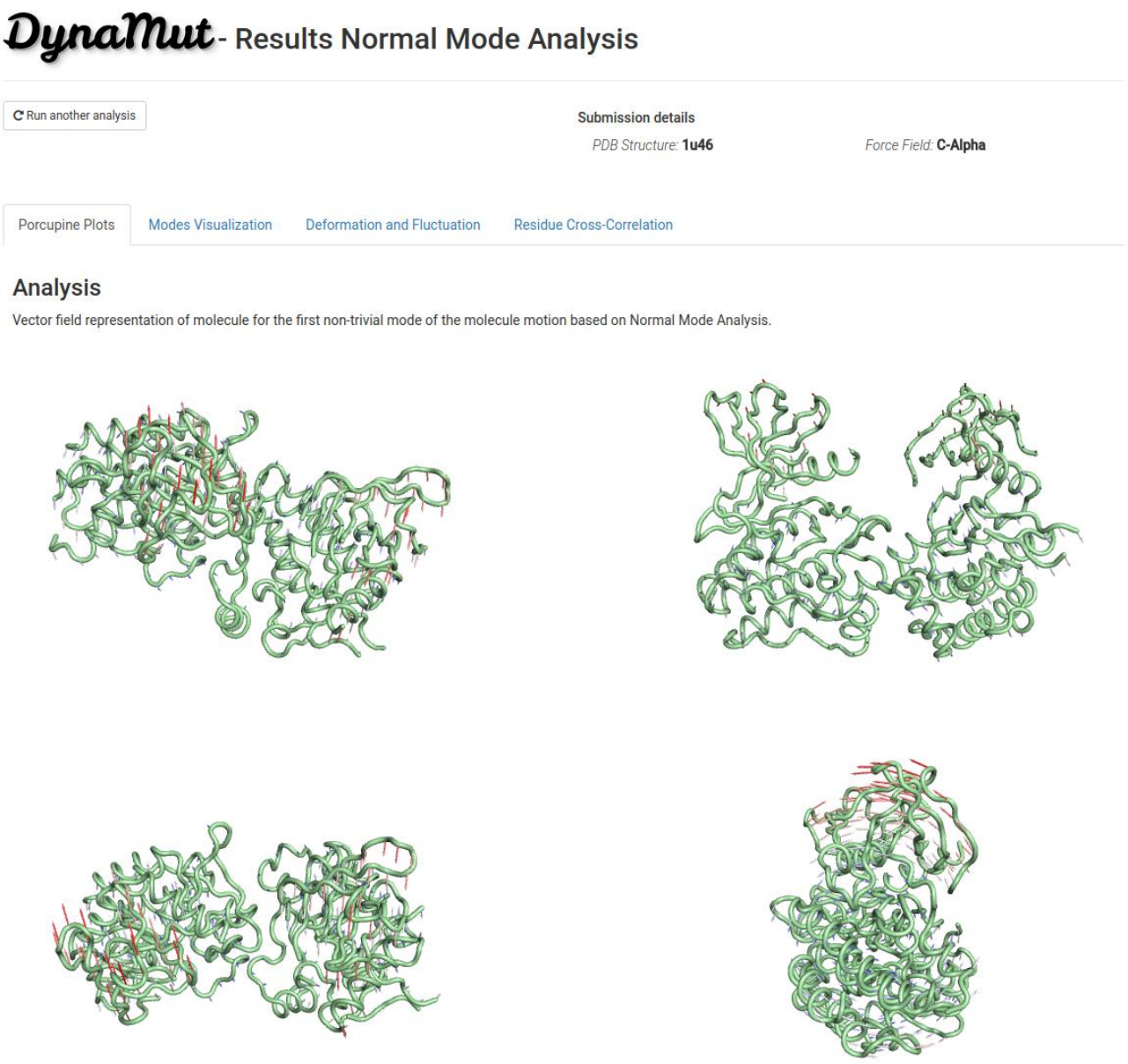


Figure S3 - Porcupine Plots on DynaMut Normal Mode Analysis output page for the example (PDB: 1U46 – Tyrosine Kinase ACK1). Protein kinase transition state from active to inactive and vice-versa requires that the protein presents a minimum flexibility no matter if the kinase is activated or not. This transitional state is directly affected by the molecule flexibility.

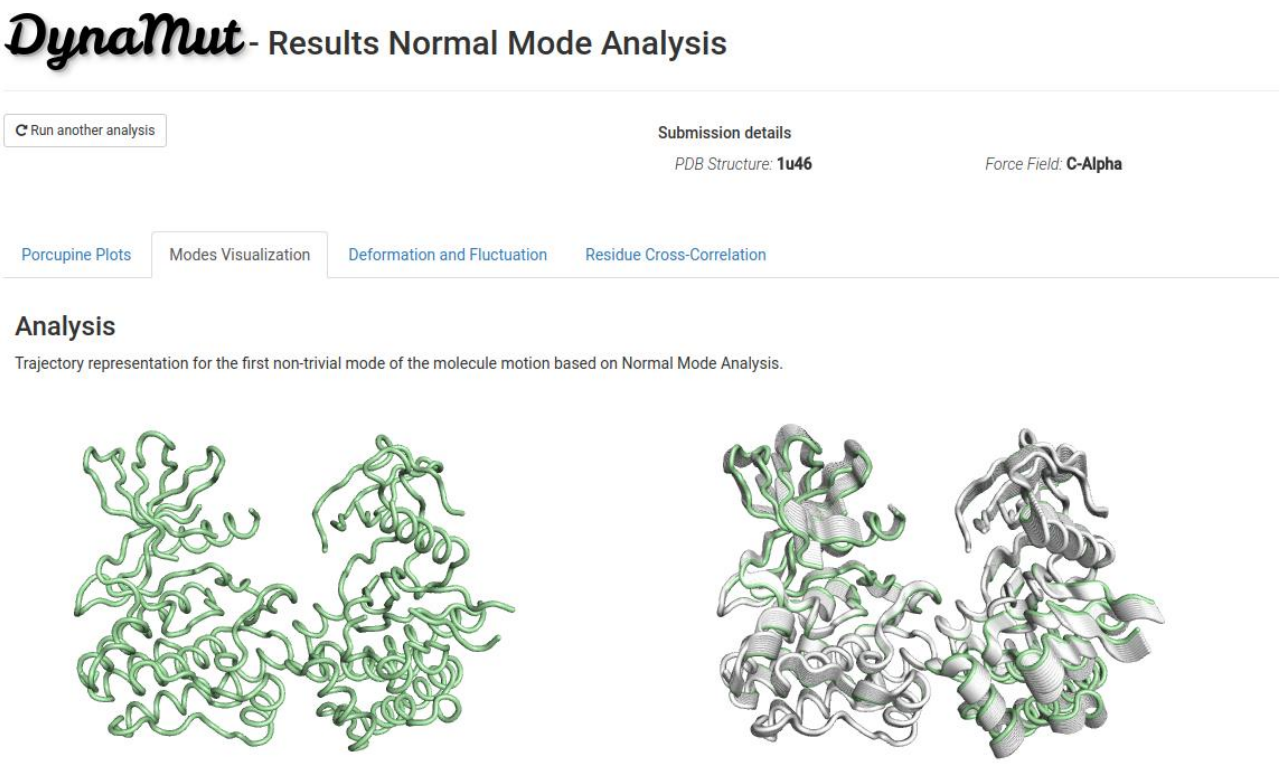


Figure S4 - Modes Visualisation on DynaMut Normal Mode Analysis output page.

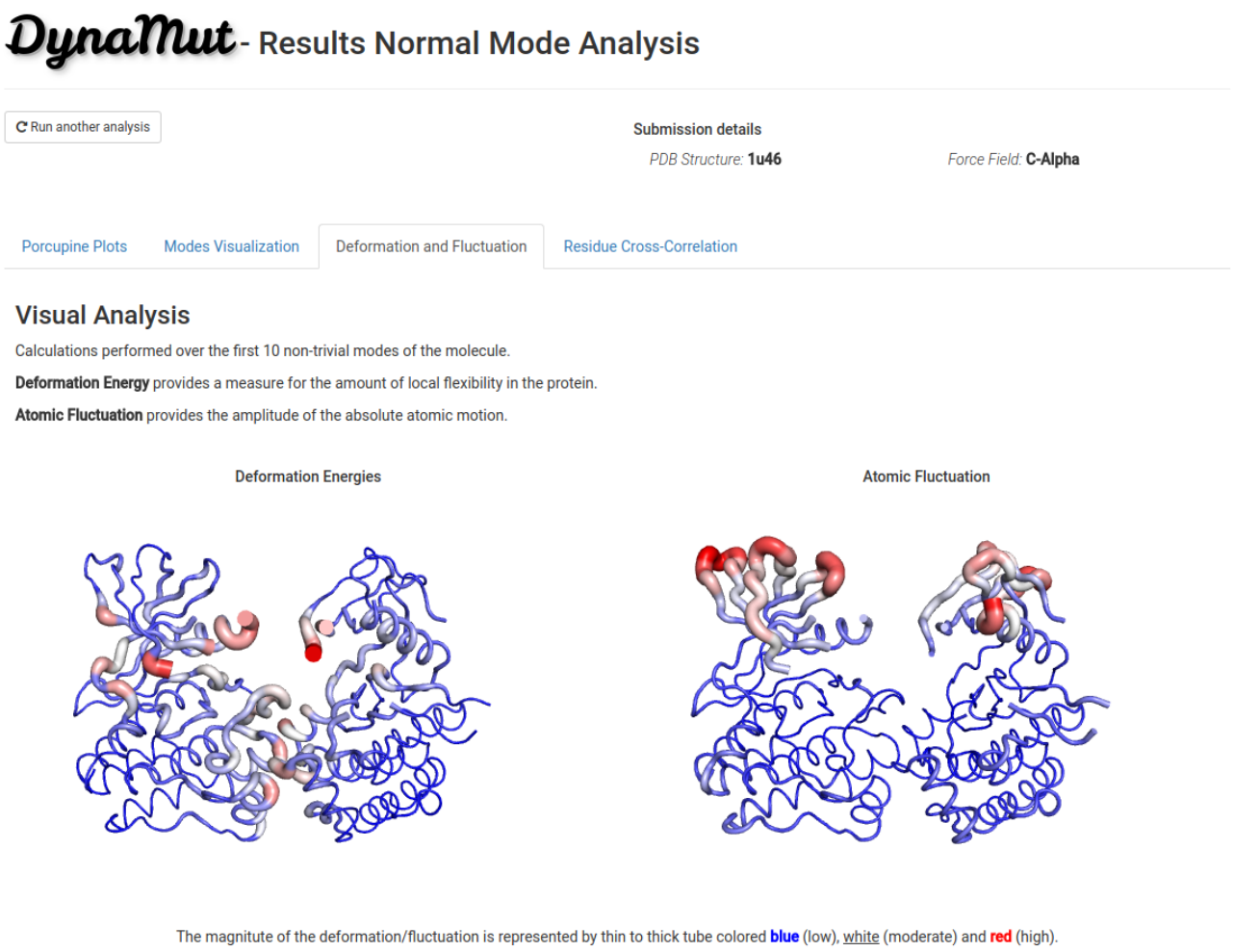


Figure S5 - Deformation energies and atomic fluctuation on DynaMut Normal Mode Analysis output page.

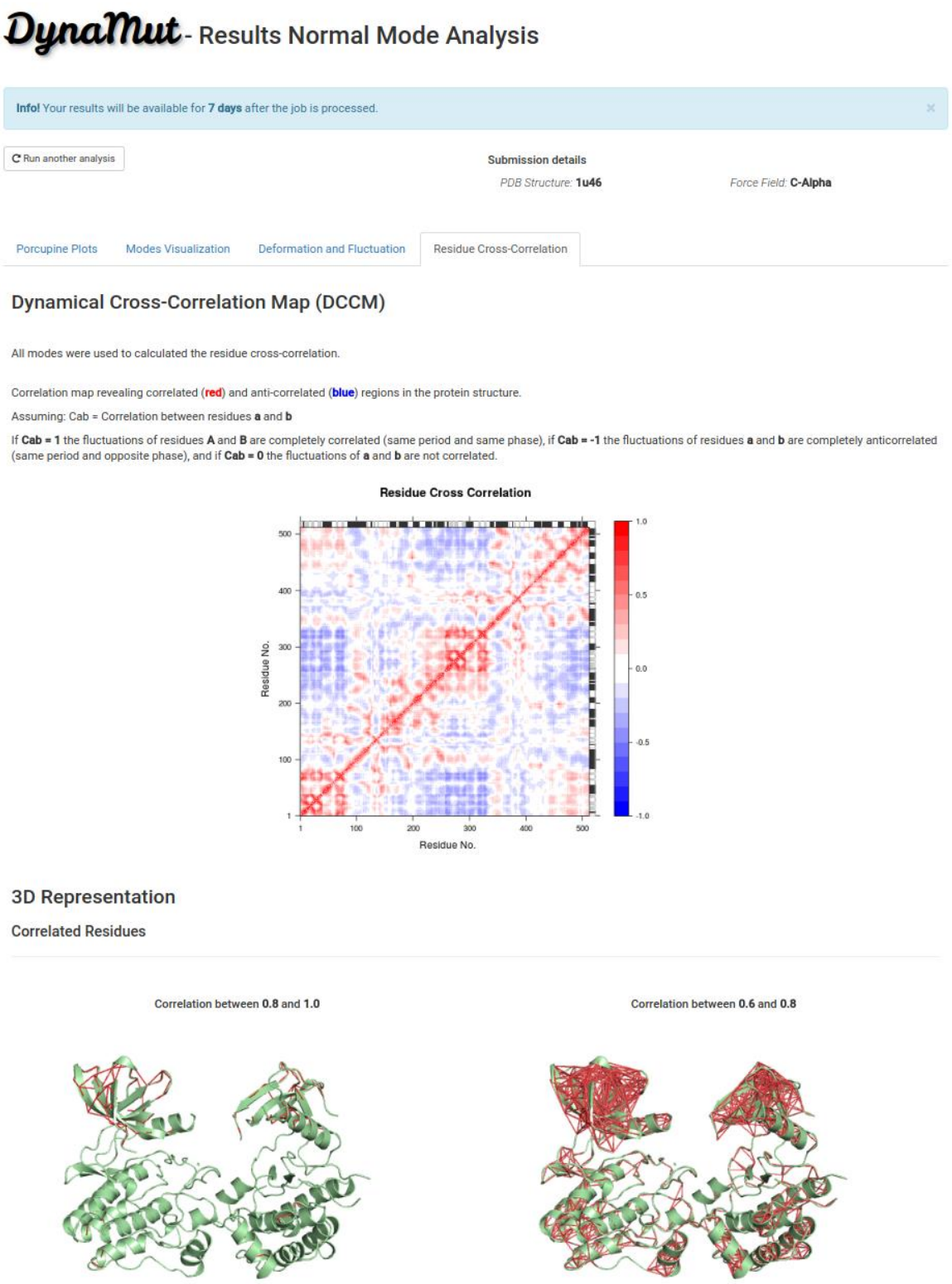


Figure S6 - Residue cross-correlation on DynaMut Normal Mode Analysis output page.

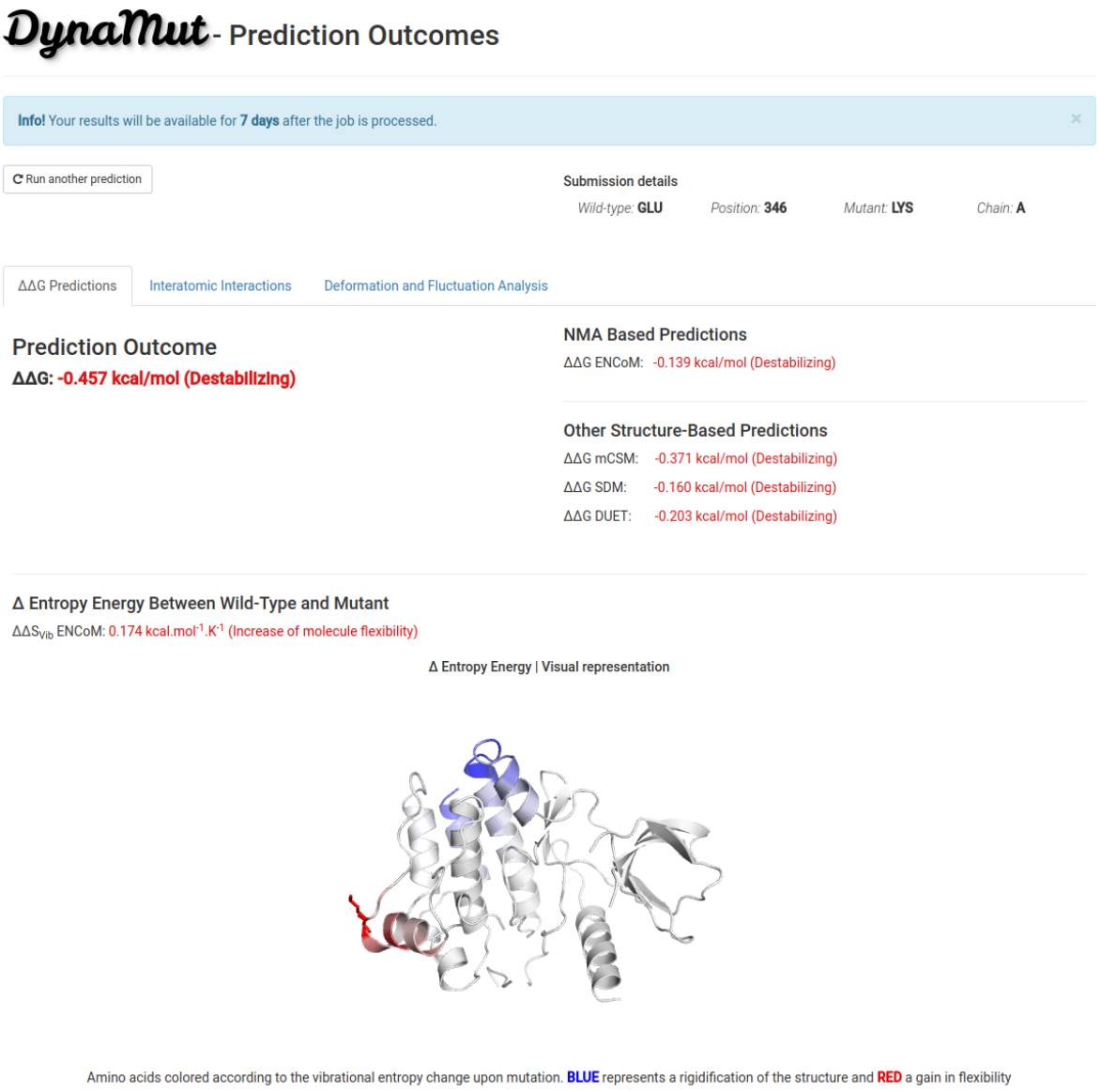


Figure S7 - Mutation effect prediction on DynaMut Prediction output page for the example (PDB: 1U46 – Tyrosine Kinase ACK1). The mechanisms by which the activating mutations affect kinases are associated with a restriction in the transition from active to inactive, resulting in one conformational state being favoured. This transitional state is directly affected by the molecule flexibility. The ΔΔG prediction outcome is shown on the top left of the page. Results for other predictive tools (NMA based and Other Structure-based approaches) are also displayed. Visual representation of the Δ Entropy Energy in which the amino acids were coloured according to the vibrational entropy change upon mutation is shown on the bottom. Blue regions indicate rigidification and red a gain in flexibility.

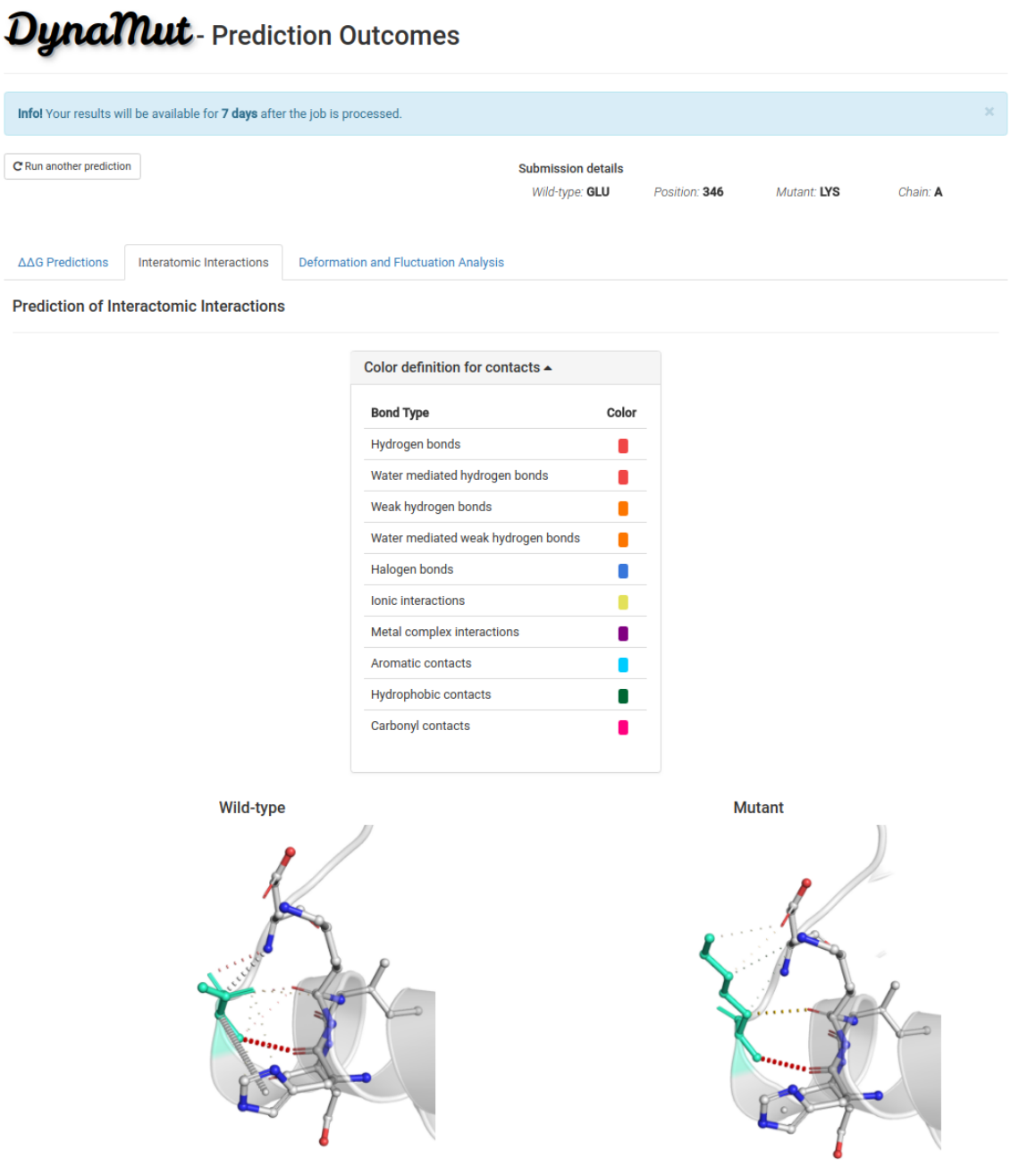


Figure S8 - Interatomic Interactions predictions of wild-type and mutant residues on output page of DynaMut. Wild-type and mutant residues are coloured in light-green and are also represented as sticks. A table with the colour definitions for each type of interaction is shown on top.

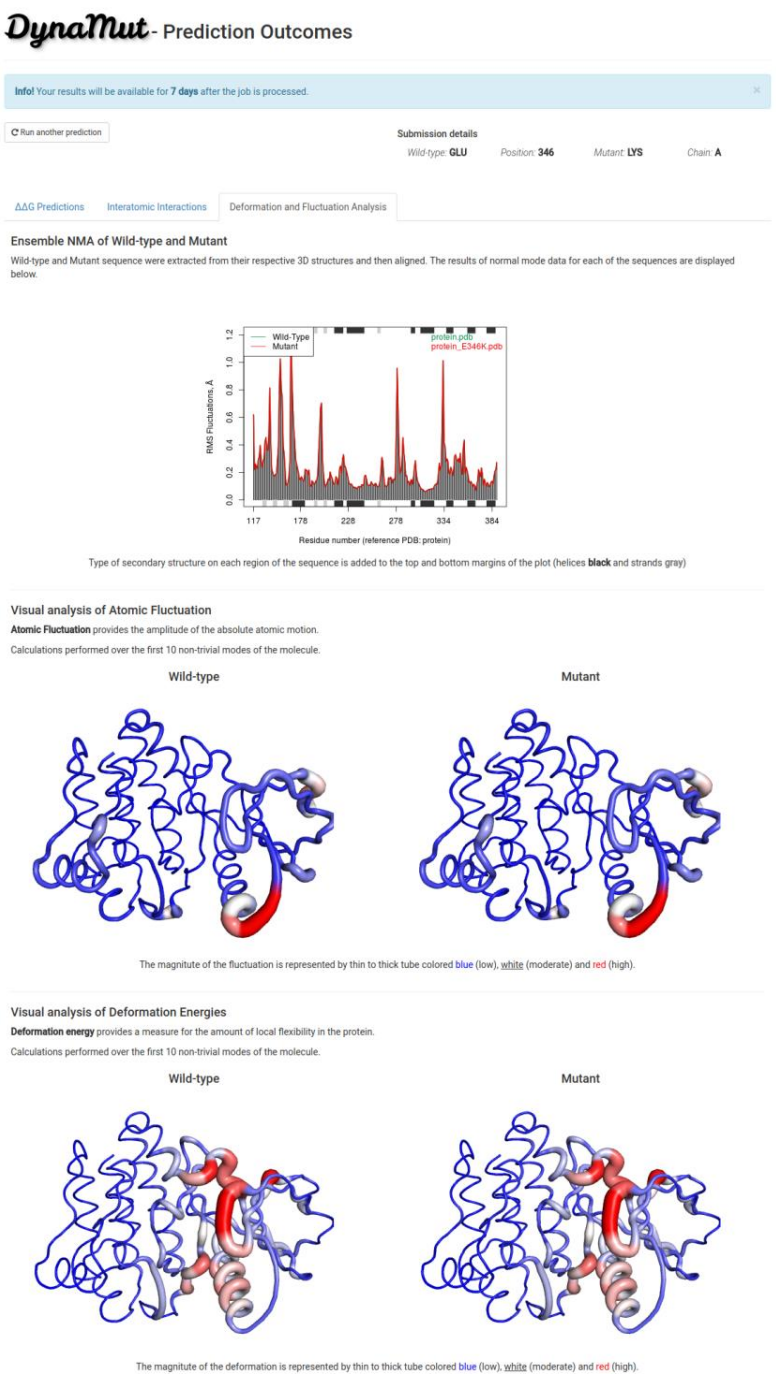


Figure S9 - Atomic fluctuation and deformation energies of the wild-type and mutant structures on output page of DynaMut. Wild-type and Mutant sequence were extracted from their respective 3D structures and then aligned. The results of normal mode data for each of the sequences are displayed on top. Visual representation of atomic fluctuation and deformation energies for wild-type (left) and mutant (right) are shown below. The magnitude of the fluctuation and deformation is represented by thin to thick tube coloured blue (low), white (moderate) and red (high).

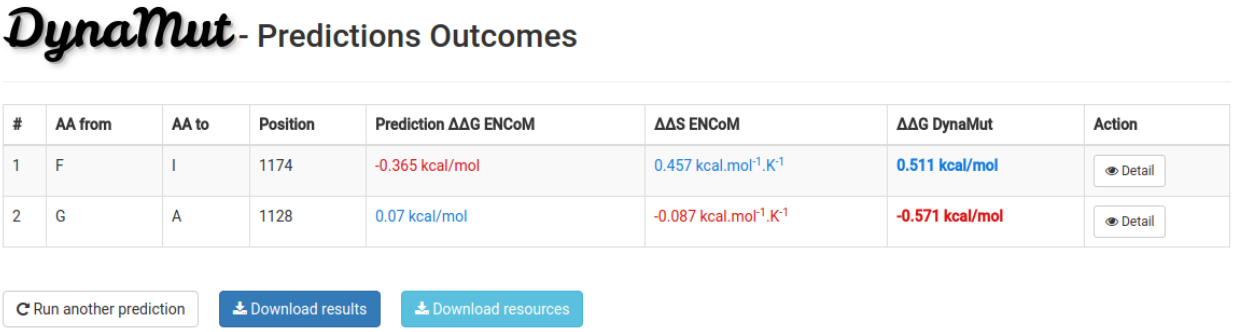


Figure S10 – Results page of DynaMut for the Mutation list option. The server output is summarised as a downloadable table, and users have the option to analyse each mutation separately, similar to the analysis of a single mutation, by clicking on the “Detail” button of each mutation on the row. All resources generated on the analysis are also available for download.

REFERENCES

1. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695-2696.
http://www.ncbi.nlm.nih.gov/pubmed/16940322
http://dx.doi.org/10.1093/bioinformatics/btl461

2. Frappier, V. and Najmanovich, R.J. (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. *PLoS Comput Biol*, **10**, e1003569.
http://www.ncbi.nlm.nih.gov/pubmed/24762569
http://dx.doi.org/10.1371/journal.pcbi.1003569

3. Karplus, M. and Kushick, J.N. (1981) Method for estimating the configurational entropy of macromolecules. *Macromolecules*, **14**, 325-332.
http://dx.doi.org/10.1021/ma50003a019

4. Pandurangan, A.P., Ochoa-Montano, B., Ascher, D.B. and Blundell, T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res*, **45**, W229-W235.
http://www.ncbi.nlm.nih.gov/pubmed/28525590
http://dx.doi.org/10.1093/nar/gkx439

5. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335-342.
http://www.ncbi.nlm.nih.gov/pubmed/24281696
http://dx.doi.org/10.1093/bioinformatics/btt691

6. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*, **42**, W314-319.
http://www.ncbi.nlm.nih.gov/pubmed/24829462
http://dx.doi.org/10.1093/nar/gku411

7. Shevade, S.K., Keerthi, S.S., Bhattacharyya, C. and Murthy, K.K. (2000) Improvements to the SMO algorithm for SVM regression. *IEEE Trans Neural Netw*, **11**, 1188-1193.
http://www.ncbi.nlm.nih.gov/pubmed/18249845
http://dx.doi.org/10.1109/72.870050

8. Sedgwick, P. (2012) Pearson's correlation coefficient. *BMJ : British Medical Journal*, **345**.
http://dx.doi.org/10.1136/bmj.e4483

9. Hyndman, R.J. and Koehler, A.B. (2006) Another look at measures of forecast accuracy. *International Journal of Forecasting*, **22**, 679-688.
http://dx.doi.org/https://doi.org/10.1016/j.ijforecast.2006.03.001

10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10-18.
http://dx.doi.org/10.1145/1656274.1656278
http://www.ncbi.nlm.nih.gov/pmc/articles/1656278

11. (2007) In Ilias, M., Kostas, K., Manolis, W. and John, S. (eds.). IOS Press.
12. Hayward, S., Kitao, A. and Go, N. (1995) Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins*, **23**, 177-186.
http://www.ncbi.nlm.nih.gov/pubmed/8592699
http://dx.doi.org/10.1002/prot.340230207

13. Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O. and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*, **80**, 505-515.
http://www.ncbi.nlm.nih.gov/pubmed/11159421
http://dx.doi.org/10.1016/S0006-3495(01)76033-X

14. Yang, L., Song, G. and Jernigan, R.L. (2009) Protein elastic network models and the ranges of cooperativity. *Proc Natl Acad Sci U S A*, **106**, 12347-12352.
http://www.ncbi.nlm.nih.gov/pubmed/19617554
http://dx.doi.org/10.1073/pnas.0902159106

15. Moritsugu, K. and Smith, J.C. (2007) Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. *Biophys J*, **93**, 3460-3469.
http://www.ncbi.nlm.nih.gov/pubmed/17693469
http://dx.doi.org/10.1529/biophysj.107.111898

16. Dehouck, Y. and Mikhailov, A.S. (2013) Effective harmonic potentials: insights into the internal cooperativity and sequence-specificity of protein dynamics. *PLoS Comput Biol*, **9**, e1003209.
http://www.ncbi.nlm.nih.gov/pubmed/24009495
http://dx.doi.org/10.1371/journal.pcbi.1003209

17. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res*, **33**, W382-388.
http://www.ncbi.nlm.nih.gov/pubmed/15980494
http://dx.doi.org/10.1093/nar/gki387

18. Capriotti, E., Fariselli, P. and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*, **33**, W306-310.
http://www.ncbi.nlm.nih.gov/pubmed/15980478
http://dx.doi.org/10.1093/nar/gki375

19. Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S. and Lackner, P. (2015) MAESTRO--multi agent stability prediction upon point mutations. *BMC Bioinformatics*, **16**, 116.
http://www.ncbi.nlm.nih.gov/pubmed/25885774
http://dx.doi.org/10.1186/s12859-015-0548-6

Received: 30 June 2020 | Revised: 27 August 2020 | Accepted: 28 August 2020

DOI: 10.1002/pro.3942

TOOLS FOR PROTEIN SCIENCE

DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations

Carlos H.M. Rodrigues[1,2] | **Douglas E.V. Pires**[1,2,3] | **David B. Ascher**[1,2,4]

[1]Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia

[2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[3]School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia

[4]Department of Biochemistry, University of Cambridge, Cambridge, UK

Correspondence
David B. Ascher and Douglas E.V. Pires, Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Victoria, Australia.
Email: david.ascher@unimelb.edu.au (D. B. A.) and douglas.pires@unimelb.edu.au (D. E.V. P.)

Funding information
Medical Research Council; National Health and Medical Research Council

Abstract

Predicting the effect of missense variations on protein stability and dynamics is important for understanding their role in diseases, and the link between protein structure and function. Approaches to estimate these changes have been proposed, but most only consider single-point missense variants and a static state of the protein, with those that incorporate dynamics are computationally expensive. Here we present DynaMut2, a web server that combines Normal Mode Analysis (NMA) methods to capture protein motion and our graph-based signatures to represent the wildtype environment to investigate the effects of single and multiple point mutations on protein stability and dynamics. DynaMut2 was able to accurately predict the effects of missense mutations on protein stability, achieving Pearson's correlation of up to 0.72 (RMSE: 1.02 kcal/mol) on a single point and 0.64 (RMSE: 1.80 kcal/mol) on multiple-point missense mutations across 10-fold cross-validation and independent blind tests. For single-point mutations, DynaMut2 achieved comparable performance with other methods when predicting variations in Gibbs Free Energy ($\Delta\Delta G$) and in melting temperature (ΔT_m). We anticipate our tool to be a valuable suite for the study of protein flexibility analysis and the study of the role of variants in disease. DynaMut2 is freely available as a web server and API at http://biosig.unimelb.edu.au/dynamut2.

KEYWORDS

dynamics, graph-based signatures, missense mutations, stability changes

1 | INTRODUCTION

Proteins are highly dynamic, metastable molecular machines. Missense mutations are associated with more than half of all known inherited diseases, however, they are often associated with more subtle molecular effects than mutations that lead to larger changes to the mature peptide. These single amino acid changes can readily disrupt the intricate network of intramolecular interactions, affecting how a protein folds, its stability, dynamics, and ultimately protein function. Beyond phenotypic outcomes,[1–22] it also has direct implications for their experimental study, protein engineering,[23,24] drug design,[25–30] and use in industrial processes.[31]

A number of approaches have been developed to predict how missense mutations affect protein stability using either sequence[32–34] or structural information.[35–37] The information from both approaches is often complementary;

however, structural methods have generally assumed a protein is static and does not consider the implications of a mutation within its conformational landscape. We previously showed that by considering both the mutation environment and the protein dynamics, we could more accurately predict the effects of single-point missense mutations.[38]

Most predictive tools, however, have been limited to single point missense variants, and the inclusion of protein dynamics computationally scales poorly with protein size. Here we present DynaMut2, an enhanced server that combines normal mode analysis with our graph-based representation of protein structure, to accurately and quickly predict the effects of single and multiple point mutations on protein stability and dynamics.

2 | RESULTS AND DISCUSSION

The DynaMut2 development workflow is summarized in Figure 1. Data on single and multiple point mutations were derived from ProTherm.[39] Given the wide range of molecular mechanisms by which mutations can impact protein function, we modeled the effects of each mutation using a range of features, including protein dynamics (NMA), wild-type residue environment, substitution propensities and contact potential scores, interatomic interactions[40] and also our well-validated graph-based signatures approach.[35,41–48] These were then used to train and test machine learning algorithms. Our predictive models were further evaluated using independent blind test sets.

2.1 | Predicting the effects of single point mutations

We initially evaluated the performance of our approach to predict changes in stability caused by single point mutations. DynaMut2 was able to achieve a Pearson's correlation of $r = 0.72$ (RMSE = 1.02 kcal/mol) for the dataset S4022, under 10-fold cross-validation, and $r = 0.68$ on S611, our non-redundant independent test set (RMSE = 1.14 kcal/mol) (Figure 2), outperforming all other methods (Table 1). The comparable performance between cross-validation and non-redundant blind test supports the generalizability of the final model. After removing 10% of outliers, performance remained

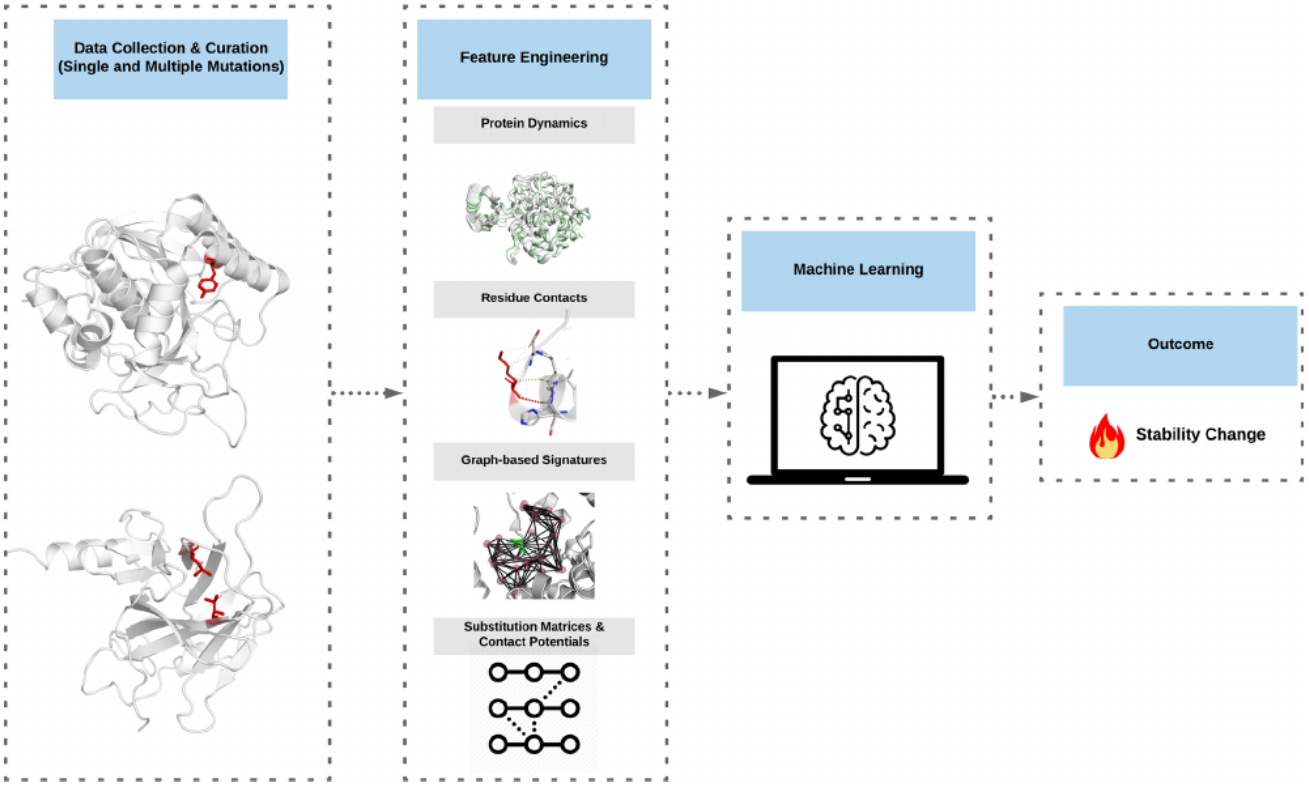


FIGURE 1 DynaMut2 workflow. The methodology for this work can be summarized into four steps: (1) data collection and curation of single and multiple mutations, (2) feature engineering to model the effects of mutations, (3) supervised machine learning, and (4) the predicted effects on stability and dynamics

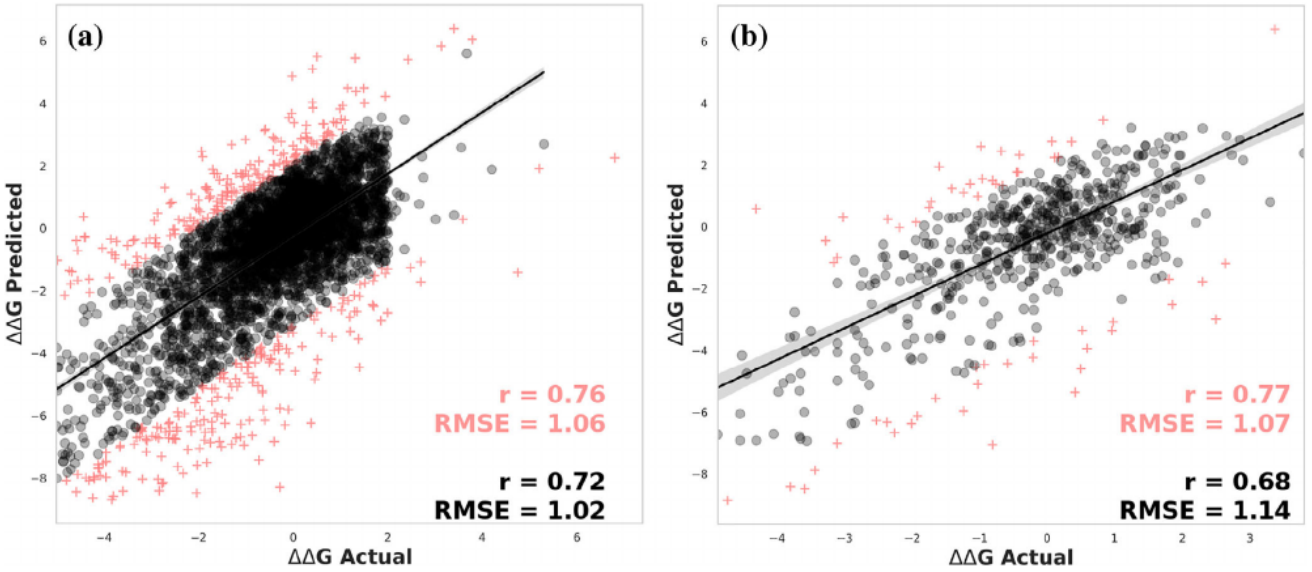


FIGURE 2 Predictive performance of DynaMut2 on 10-fold cross-validation (a) and non-redundant test sets (b) for single point mutations. 10% of outliers are shown as pink crosses

TABLE 1 Comparative performance across the non-redundant test set S611

	Overall		Stabilizing mutations		Destabilizing mutations		
Method	**Pearson (*r*)**	**RMSE (kcal/Mol)**	**Pearson (*r*)**	**RMSE (kcal/Mol)**	**Pearson (*r*)**	**RMSE (kcal/Mol)**	**AUC**
DynaMut2	**0.68**	**1.14**	**0.51**	**1.02**	**0.62**	**0.91**	**0.68**
DynaMut1	0.49*	1.38+	0.47	1.24	0.55	1.01	0.62
SDM	0.35*	1.93+	0.15*	2.00+	0.36*	1.86+	0.60#
mCSM	0.46*	1.42+	0.11*	1.81+	0.56	0.98	0.56#
DUET	0.48*	1.40+	0.09*	1.75+	0.58	1.00	0.56#
ENCoM	−0.14*	2.03+	−0.01*	1.94+	−0.18*	2.09+	0.41#
Maestro	−0.36*	1.55+	0.27*	1.17	0.43*	1.81+	0.46#
I-mutant	0.33*	1.47+	0.03*	1.83+	0.49*	1.09+	0.51#
MUpro[a]	0.15*	1.71+	−0.05*	2.15+	0.23*	1.21+	0.50#

*p Value < .05 compared with DynaMut2 using z test.
#p Value < .05 compared with DynaMut2 using t test.
+p Value < .05 compared with DynaMut2 using Diebold-Mariano test.
[a]48 mutations were left out due to input issues.

consistent for the training set with $r = 0.76$ (RMSE = 1.06), and increased to $r = 0.77$ (RMSE = 1.07) on the test set. No significant differences in the distributions of properties were observed for the outliers compared to the overall dataset.

Due to the natural imbalance between stabilizing and destabilizing mutations present in the training and evaluation data (Figure S1), we further analyzed performance on the respective classes separately. Across the non-redundant validation se S611, DynaMut2 achieved a Pearson's correlation of $r = 0.62$ (RMSE = 1.75) and $r = 0.51$ (RMSE = 1.88) on destabilizing and stabilizing mutations, respectively. The slightly lower performance toward stabilizing mutations was expected due to the imbalanced distribution of data but was significantly improved compared to previous methods (Table 1). These results remained consistent when we compared the performances using rank coefficient scores Kendall and Spearman (Table S1). This was further reflected in the ability of DynaMut2 to correctly classify stabilizing and destabilizing mutations (AUC 0.68), outperforming previous approaches.

To further investigate potential biases in the predictive performance, we evaluated the performance of

DynaMut2 on subsets derived from the O2567 dataset.[49] DynaMut2 showed significantly better performance than all other approaches for mutations on buried residues (RSA ≤30%; Table S2). A small deterioration in performance is observed on mutations on exposed residues (RSA > 30%; Table S3), likely to be related to the smaller number features captured by the graph-based signatures in DynaMut2; however, our method still achieved comparable results to mCSM, MAESTRO and SDM, and outperformed other approaches. Evaluating the performance on different protein CATH classifications, DynaMut2 outperformed other approaches across β-sheet structures (Table S4), and α-helix and β-helix structures (Table S5). The size of the protein being mutated did not affect performance, with comparable performance between larger proteins (>150 residues; Table S6) and small proteins (<150 residues; Table S7), outperforming all other evaluated approaches. Similarly, DynaMut2 performance was similar to mutations from large to small residues (Table S8), from small to large residues (Table S9), or for mutations between residues of comparable sizes (Table S10). Encouragingly, DynaMut2 outperformed all other approaches on mutations leading to a change in volume and demonstrated comparable performance to the top approaches for mutations between residues of similar volume. Overall, this highlighted that DynaMut2 predictive performance across all single-point mutations was significantly more balanced and less biased than all other methods evaluated.

We further evaluated the performance of our model across an additional independent test set, S276. DynaMut2 achieved a Pearson's correlation of 0.52, comparable with the best-performing methods (Table 2) and significantly better than MUpro.[50] Although not directly comparable, as there is a correlation between changes upon mutation in stability (ΔG) and thermal stability (T_m),[51] the performance of DynaMut2 on predicting changes in melting temperature was assessed using the blind test set S173. Results were stratified by protein structure and summarized in Table S11. Overall, DynaMut2 ranks fourth among the methods evaluated; however, performances of all methods varied greatly between structures. These results indicate a possible challenge in accurately predicting the thermal stability effects of mutations on a more diverse set of proteins.

TABLE 2 Comparative performance across the S276 blind test of experimental $\Delta\Delta G$

Method	*R*	MAE (kcal/Mol)
DynaMut2	0.52	0.88
DeepDDG	0.55	0.86
SDM	0.48	1.02
mCSM	0.46	0.90
I-mutant	0.45	0.91
STRUM	0.44	0.88
MUpro	0.19*	1.06

*p Value < .05 compared with DynaMut2 using z test.

2.2 | Predicting the effects of multiple point mutations

The performance of our approach to predict the effects of multiple point mutations on protein stability was then assessed. DynaMut2 achieved a Pearson's correlation of $r = 0.71$ (RMSE = 1.66 kcal/mol) under 10-fold cross-validation and $r = 0.67$ (RMSE = 1.79 kcal/mol) on our non-redundant test set. The comparable performance between cross-validation and blind test set again gave confidence in the generalizability of the approach. This significantly outperformed the previously reported performances of DDGun, DDGun3D, Maestro, and FoldX, whose correlations ranged from 0.37 to 0.55 on the experimental multiple point mutations in ProTherm.[52] Performances were consistent when considering only 90% of the data, with DynaMut2 achieving $r = 0.82$ (RMSE = 1.91) and $r = 0.80$ (RMSE = 2.01) on 10-fold cross-validation and blind-test, respectively (Figure 3). This indicates that outlier predictions were not having a significant effect on the correlations.

The unbalanced nature of the training dataset was evident when we analyzed the performance of our final model on stabilizing and destabilizing multiple mutations separately (Table 3). Overall, DynaMut2 was able to correctly classify 80% of multiple missense mutations (AUC 0.84) in the blind test set, including 93% of the destabilizing mutations, providing confidence in the ranking ability of the approach. As expected, however, across our non-redundant test set DynaMut2 shows a better performance toward predicting multiple mutations with a destabilizing effect, achieving a Pearson's correlation of $r = 0.56$ (RMSE = 2.66), while for stabilizing entries the performance drops to $r = 0.42$ (RMSE = 2.94). These results indicate a need for new experimental data on multiple point mutations, especially those with stabilizing effects, since the use of hypothetical reverse mutations is likely to add more uncertainty to the model.

2.3 | Web server

We have implemented DynaMut2 as a freely available and user-friendly web server (http://biosig.unimelb.edu.au/

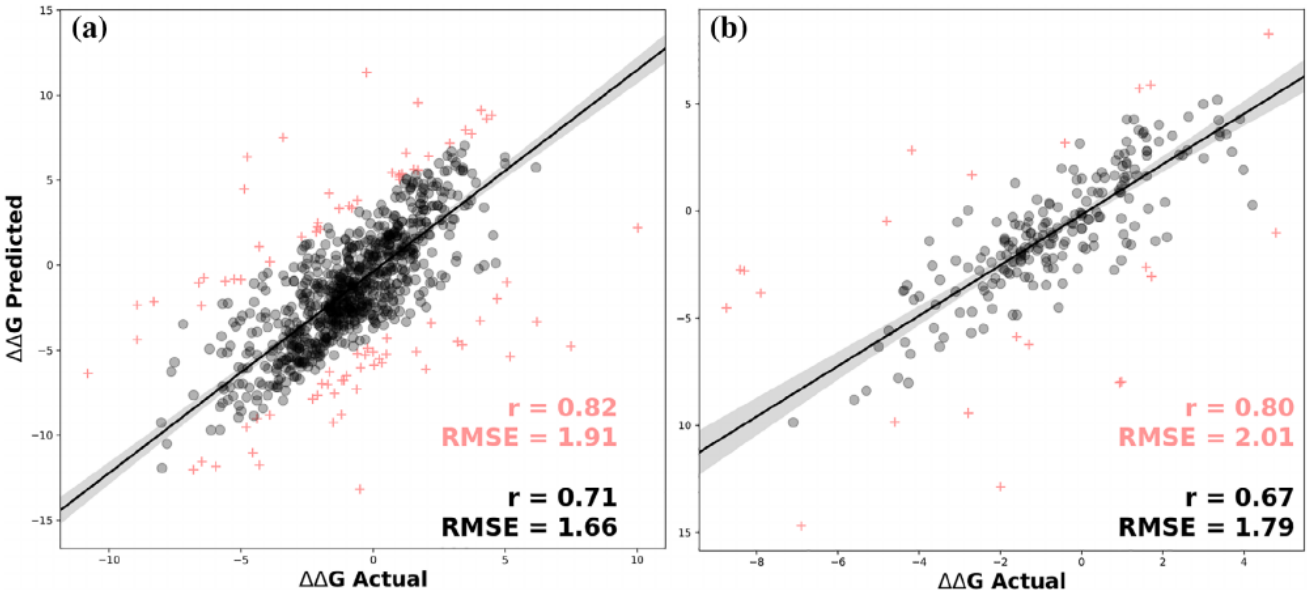


FIGURE 3 Predictive performance of DynaMut2 on 10-fold cross-validation (a) and non-redundant test sets (b) for multiple point mutations. 10% of outliers are shown as pink crosses

TABLE 3 Comparative performance on multiple mutations prediction across different correlation coefficients

	Overall			Stabilizing			Destabilizing		
Methods	r_p	Tau	r_s	r_p	Tau	r_s	r_p	Tau	r_s
DynaMut2	**0.71**	**0.58**	**0.75**	**0.42**	**0.38**	**0.53**	**0.56**	**0.47**	**0.63**
MAESTRO	0.19*	0.13+	0.19#	0.12*	0.07+	0.08#	0.21*	0.14+	0.21#
FoldX	0.33*	0.21+	0.31#	0.04*	0.06+	0.09#	0.30*	0.19+	0.27#

*p Value < .05 compared with DynaMut2 using Fisher r-to-z transformation.
+p Value <.05 by transforming tau-to-r followed by Fisher r-to-z transformation.
#p Value <.05 by transforming rho-to-r followed by Fisher r-to-z transformation.

dynamut2/). The frontend was developed using Materializecss version 1.0.0 and the backend uses the Flask module (1.0.2) from the Python programming language. The web server is hosted on a Linux machine running Apache.

2.4 | Input

DynaMut2 can be used in three different ways[1]: predicting $\Delta\Delta G$ for single point mutations,[2] predicting $\Delta\Delta G$ for multiple point mutations (up to three), and also analysis of protein dynamics based on NMA. For predicting single point mutations, similarly to our previous implementation of DynaMut, two different inputs are available: "Single mutation" and "List of Mutations". For the Single Mutation option, users are required to provide a protein structure on PDB format or provide a four-digit code of an entry on the PDB, the chain identifier where the mutation occurs and the point mutation defined as string comprising wild-type residue one-letter code, residue position, and mutant residue one-letter code. For the List of Mutations option, users must provide the structure of the protein, similarly to the Single Mutation option, and also upload a file with the list of variants (one per line), following the same mutation code previously defined.

For predicting the effects of multiple mutations, users are required to provide the structure of the protein, as previously described, and also the multiple mutations separated by a comma. DynaMut2 also allows for submitting a list of multiple point mutations to be analyzed in batch. These can be input by uploading a file with one entry of multiple mutations separated by comma per line.

Alternatively, for protein dynamic analysis, users are required to input the protein structure by uploading a file using the PDB format or provide a valid four-digit code for a PDB entry, and also select one of the force fields available to guide structural interactions for NMA. All force field options available are detailed in Table S12.

2.5 | Output

For single point mutations on the "Single Mutation" option, predicted $\Delta\Delta G$ is shown at the top with details of users' input and also the wild-type residue environment (Figure 4). All interatomic contacts calculated with Arpeggio are also displayed as an interactive viewer using NGL viewer.[53] On the "Mutation List" option, the results are displayed as a downloadable table with options to view details of each variation separately, similarly to the analysis provided by the "Single Mutation", option.

The results for multiple mutations, predictions are displayed at the top of the page with detailed information for each mutation, if a list is provided these results are

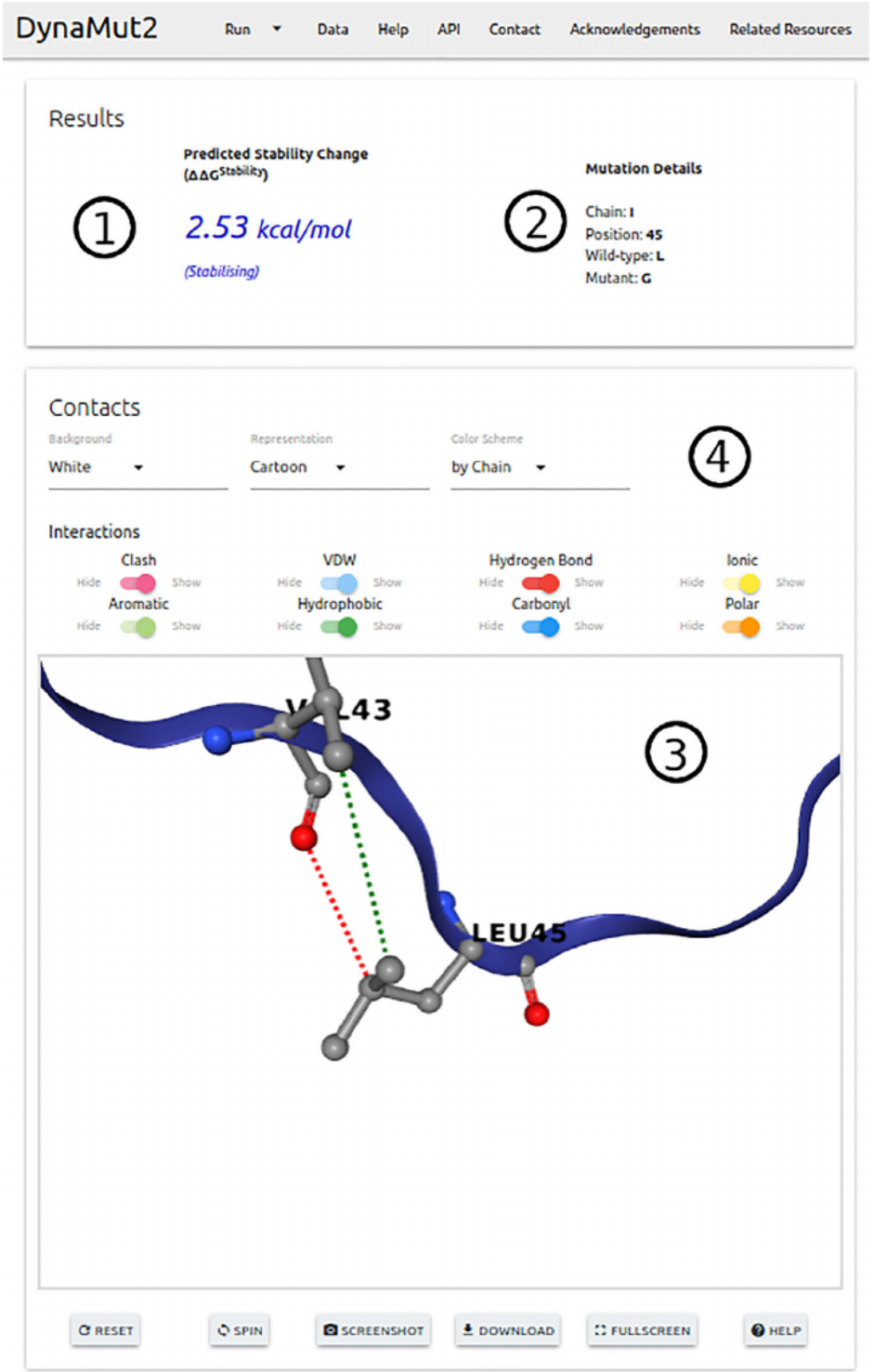


FIGURE 4 DynaMut2 results page. The figure depicts the prediction results page for single-point mutations. The predicted effect of a mutation in stability and dynamics is given as the variation in Gibbs Free Energy (in Kcal/mol) (1), together with complementary information about the mutation provided (2). Users can inspect the wild-type residue environment via an interactive viewer (3), which also allows the visualization of non-covalent interactions established by the mutated residue

shown as a table. An interactive viewer allowing for the analysis of residue contacts is also available.

For NMA submissions, the results are displayed on three panels. The first two provide information on trajectory representation of the molecule motion and porcupine plots, summarizing vector field representation, for the first non-trivial modes. Finally, the last panel displays a residue correlation matrix and structural representations using all modes.

2.6 | API

DynaMut2 conveniently offers an API (Application Programming Interface) to assist users in integrating our predictive tool into their research pipelines. All jobs submitted to DynaMut2 are labeled with a unique identifier, which is used to query the status of the job. Input fields follow the same rules of our website implementation. A full description of these with examples using curl and Python are available at http://biosig.unimelb.edu.au/dynamut2/api.

2.7 | Processing time

Finally, we compared the performance of DynaMut2 with our previous implementation, DynaMut, in terms of processing time for single-point mutations on six different protein structures. For each structure, we submitted a single-point mutation to each server and computed the processing time in seconds. This procedure was repeated 10 times for each mutation and the results are summarized in Table S13 and Figure S2. Clearly, the greater the number of residues comprising the protein structure the longer it takes for our NMA based methods to generate predictions. However, DynaMut2 runs much faster than DynaMut on all sets of experiments with very little differences in each repetition and an improvement of more than six times in the worst-case scenario.

3 | CONCLUSION

Here we present DynaMut2, a tool that incorporates information on protein dynamics and structural environment properties of wild-type residue with our graph-based signatures approach to provide an accurate prediction of mutation effects on stability and dynamics for single and multiple point mutations. Our updated server has shown to outperform other methods on predicting changes in stability caused by single point mutations and also comparable results for when used for estimating ΔT_m. In addition, our new approach was significantly faster compared to the original DynaMut, which will be of great benefit toward large scale analysis and large structures. Finally, we have extended our method to predict the effects of multiple point mutations (double and triple mutants) and an API, which conveniently enables users to programmatically run predictions and represents a great contribution in terms of a novelty for this type of tool. We believe DynaMut2 represents an invaluable resource for the study of protein dynamics and to help understand the role of mutations in diseases. Web server and API with examples are freely available at http://biosig.unimelb.edu.au/dynamut2.

4 | MATERIALS AND METHODS

4.1 | Data set

We have collected experimental data on 2,648 single point mutations on 125 globular proteins from Protherm.[39] Of these, 2,080 are destabilizing ($\Delta\Delta G < 0.0$ kcal/mol) and 568 stabilizing ($\Delta\Delta G > 0.0$ kcal/mol) (Figure S1). To minimize the imbalanced nature of our dataset (Figure S1) and as a sanity check evidenced by other studies, here we use hypothetical reverse mutations.[36,54] However, differently from our previous implementation of DynaMut, hypothetical reverse mutations with more drastic changes on Gibbs free energy ($\Delta\Delta G < -2.0$ kcal/mol or $\Delta\Delta G > 2.0$ kcal/mol) were left out of our study due to uncertainties about the quality and biological significance of the modeled mutant. Our final dataset comprised 4,633 mutations (2,640 destabilizing and 1,993 stabilizing), which were split into 4,022 entries (S4022) for training our predictive model and a non-redundant test set comprising 611 entries (S611), following the protocol from our initial version of DynaMut.[38] For further performance evaluation and comparison with other methods, here we also consider a test set of 276 mutations (S276) with low sequence identity to proteins in the original ProTherm dataset, and an independent test set comprising 173 variants (S173) in six proteins with experimental melting temperature changes available (ΔT_m). The latter includes the structure of guanylate kinase (GK) obtained through homology modeling with Modeller[55] using the mouse GK as a template (PDB: 1LVG), similarly to previous works.[56,57]

For the data on multiple point mutations, we were able to extract 1,323 entries from ProTherm; however, since the majority of entries were double and triple mutants (Figure S3) and for the sake of simplicity, here we only considered those two types. Our final dataset comprised 1,098 entries (710 destabilizing and

388 stabilizing) (Figure S4), which were randomly split into train and test sets comprising 872 and 227 entries, respectively.

In this study, we prioritize the use of biological assembly structures author assigned, if not available, for structures generated using NMR for instance, the asymmetric unit was considered. All data used in this study is freely available for download at http://biosig.unimelb.edu.au/dynamut2/data.

4.2 | Normal model analysis

NMA provides a valuable approach for the study of dynamics and accessible conformations in a system as an alternative to time-consuming and computationally expensive Molecular Dynamics simulations. Similarly with our previous work, here we incorporated dynamics properties extracted from the protein structure generated with the module NMA of the bio3D tool.[58]

4.3 | Graph-based signatures

Our in-house graph-based signatures approach to represent molecular structures[35,59–61] has proven to be successful for a range of applications toward the study of protein structure and changes carried out by missense mutations,[35,37,41–48] including phenotypic changes.[16,62,63] These signatures comprise physicochemical and geometrical properties from the wild-type environment based on distance patterns mined from the 3D structure by representing atoms as nodes and their interactions as edges. Physicochemical properties are then defined based upon the amino acid properties, namely pharmacophore, and distance patterns between atoms are summarized as cumulative distribution functions.

4.4 | Analysis of mutation effects

Changes in Gibbs Free energy of folding can occur due to a myriad of factors related and in order to incorporate these properties, we used Arpeggio[40] to calculate the number of hydrophobic contacts involving the wild-type residue and contact potential scores from AAINDEX database.[64]

4.5 | Machine learning

In this study, we used the implementation of the Random Forest algorithm available on the scikit-learn Python library for both the prediction of $\Delta\Delta G$ for single and multiple mutations. In order to avoid the curse of dimensionality and improve performance, we selected our features using an incremental stepwise greedy approach.

5 | GENERAL STATEMENT

Small changes in proteins can have large phenotypic outcomes. By considering the changes of mutations within the context of the protein 3D structure, we have been able to accurately predict the molecular consequences of single and multiple point mutations on protein folding, stability, and dynamics. We have made this tool available through an easy to use website and API.

ACKNOWLEDGMENTS

We wish to thank Lim Yong Shan and Chandra Verma, Bioinformatics Institute, A*STAR for their help in testing and evaluating the webserver. This work thanks, the funding of Melbourne Research Scholarships (to C.H.M. R); Medical Research Council (MRC) (MR/M026302/1 to D.B.A, D.E.V.P); National Health and Medical Research Council of Australia (GNT1174405 to D.B.A.); Jack Brockhoff Foundation (JBF 4186, 2016 to D.B.A.); Wellcome Trust (200814/Z/16/Z); Supported in part by the State Government of Victoria's OIS Program.

AUTHOR CONTRIBUTIONS

Carlos Rodrigues: Data curation; formal analysis; investigation; methodology; validation; visualization; writing-original draft. **Douglas Pires:** Formal analysis; methodology; writing-review and editing. **David Ascher:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; supervision; validation; writing-original draft; writing-review and editing.

CONFLICT OF INTEREST

No conflict of interest declared.

ORCID

David B. Ascher https://orcid.org/0000-0003-2948-2413

REFERENCES

1. Byrne JA, Strautnieks SS, Ihrke G, et al. Missense mutations and single nucleotide polymorphisms in ABCB11 impair bile salt export pump processing and function or disrupt pre-messenger RNA splicing. Hepatology. 2009;49:553–567.
2. Pires DE, Chen J, Blundell TL, Ascher DB. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep. 2016;6:19848.

3. Jafri M, Wake NC, Ascher DB, et al. Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov. 2015;5:723–729.
4. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR. Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: Identification of novel mutations. JIMD Rep. 2015;24:3–11.
5. Nemethova M, Radvanszky J, Kadasi L, et al. Twelve novel HGD gene variants identified in 99 alkaptonuria patients: Focus on 'black bone disease' in Italy. Eur J Hum Genet. 2016; 24:66–72.
6. Casey RT, Ascher DB, Rattenberry E, et al. SDHA related tumorigenesis: A new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med. 2017;5:237–250.
7. Soardi FC, Machado-Silva A, Linhares ND, et al. Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med. 2017;2:7.
8. Hawkey J, Ascher DB, Judd LM, et al. Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom. 2018;4:e000165.
9. Hnizda A, Fabry M, Moriyama T, et al. Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia. 2018;32: 1393–1403.
10. Holt KE, McAdam P, Thai PVK, et al. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet. 2018;50:849–856.
11. Karmakar M, Globan M, Fyfe JAM, et al. Analysis of a novel pncA mutation for susceptibility to pyrazinamide yherapy. Am J Respir Crit Care Med. 2018;198:541–544.
12. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N. Understanding molecular consequences of putative drug resistant mutations in *Mycobacterium tuberculosis*. Sci Rep. 2018;8: 15356.
13. Vedithi SC, Malhotra S, Das M, et al. Structural implications of mutations conferring rifampin resistance in mycobacterium leprae. Sci Rep. 2018;8:5016.
14. Ascher DB, Spiga O, Sekelska M, et al. Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. Eur J Hum Genet. 2019;27:888–902.
15. Hildebrand JM, Kauppi M, Majewski IJ, et al. A missense mutation in the MLKL brace region promotes lethal neonatal inflammation and hematopoietic dysfunction. Nat Commun. 2020;11:3150.
16. Karmakar M, Rodrigues CHM, Horan K, Denholm JT, Ascher DB. Structure guided prediction of pyrazinamide resistance mutations in pncA. Sci Rep. 2020;10:1875.
17. Vedithi SC, Rodrigues CHM, Portelli S, et al. Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of RNA polymerase in mycobacterium leprae. Comput Struct Biotechnol J. 2020;18: 271–286.
18. Andrews KA, Ascher DB, Pires DEV, et al. Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet. 2018;55:384–394.
19. Trezza A, Bernini A, Langella A, et al. A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci. 2017;58:5320–5328.
20. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW. Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep. 2014;4:4765.
21. Phelan J, Coll F, McNerney R, et al. *Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med. 2016;14:31.
22. Silvino AC, Costa GL, Araujo FC, et al. Variation in human cytochrome P-450 drug-metabolism genes: A gateway to the understanding of *Plasmodium vivax* relapses. PLoS One. 2016; 11:e0160172.
23. Raghunathan G, Sokalingam S, Soundrarajan N, Madan B, Munussami G, Lee SG. Modulation of protein stability and aggregation properties by surface charge engineering. Mol Biosyst. 2013;9:2379–2389.
24. Coelho MB, Ascher DB, Gooding C, et al. Functional interactions between polypyrimidine tract binding protein and PRI peptide ligand containing proteins. Biochem Soc Trans. 2016; 44:1058–1065.
25. Karpiyevich M, Adjalley S, Mol M, et al. Nedd8 hydrolysis by UCH proteases in plasmodium parasites. PLoS Pathog. 2019;15: e1008086.
26. Singh V, Donini S, Pacitto A, et al. The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis. 2017;3:5–17.
27. Singh V, Pacitto A, Donini S, et al. Synthesis and structure-activity relationship of 1-(5-isoquinolinesulfonyl)piperazine analogues as inhibitors of *Mycobacterium tuberculosis* IMPDH. Eur J Med Chem. 2019;174:309–329.
28. White RR, Ponsford AH, Weekes MP, et al. Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog. 2016;12:e1005977.
29. Park Y, Pacitto A, Bayliss T, et al. Essential but not vulnerable: Indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against *Mycobacterium tuberculosis*. ACS Infect Dis. 2017;3:18–33.
30. Portelli S, Olshansky M, Rodrigues CH, et al. Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. Nat Genet. 2020; in press.
31. Kim DH, Kim MS. Hydrogenases for biological hydrogen production. Bioresour Technol. 2011;102:8423–8431.
32. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33:W306–W310.
33. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–3814.
34. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–249.
35. Pires DE, Ascher DB, Blundell TL. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30:335–342.
36. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL. SDM: A server for predicting effects of mutations on protein stability. Nucleic Acids Res. 2017;45:W229–W235.

37. Pires DE, Ascher DB, Blundell TL. DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014; 42:W314–W319.
38. Rodrigues CH, Pires DE, Ascher DB. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. 2018;46:W350–W355.
39. Kumar MD, Bava KA, Gromiha MM, et al. ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34:D204–D206.
40. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol. 2017;429:365–371.
41. Pires DE, Ascher DB. mCSM-AB: A web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res. 2016;44:W469–W473.
42. Pires DE, Ascher DB. CSM-lig: A web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res. 2016;44:W557–W561.
43. Pires DE, Blundell TL, Ascher DB. mCSM-lig: Quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep. 2016;6:29575.
44. Pires DEV, Ascher DB. mCSM-NA: Predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res. 2017;45:W241–W246.
45. Myung Y, Pires DEV, Ascher DB. mmCSM-AB: Guiding rational antibody engineering through multiple point mutations. Nucleic Acids Res. 2020;48:W125–W131.
46. Myung Y, Rodrigues CHM, Ascher DB, Pires DEV. mCSM-AB2: Guiding rational antibody design using graph-based signatures. Bioinformatics. 2020;36:1453–1459.
47. Pires DEV, Rodrigues CHM, Ascher DB. mCSM-membrane: Predicting the effects of mutations on transmembrane proteins. Nucleic Acids Res. 2020;48:W147–W153.
48. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: Predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res. 2019;47:W338–W344.
49. Caldararu O, Mehra R, Blundell TL, Kepp KP. Systematic investigation of the data set dependency of protein stability predictors. J Chem Inf Model. 2020. https://doi.org/10.1021/acs.jcim.0c00591.
50. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006;62:1125–1132.
51. Watson MD, Monroe J, Raleigh DP. Size-dependent relationships between protein stability and thermal unfolding temperature have important implications for analysis of protein energetics and high-throughput assays of protein-ligand interactions. J Phys Chem B. 2018;122:5278–5285.
52. Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P. DDGun: An untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinformatics. 2019;20:335.
53. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: Web-based molecular graphics for large complexes. Bioinformatics. 2018;34:3755–3758.
54. Thiltgen G, Goldstein RA. Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One. 2012;7:e46084.
55. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234:779–815.
56. Sekulic N, Shuvalova L, Spangenberg O, Konrad M, Lavie A. Structural characterization of the closed conformation of mouse guanylate kinase. J Biol Chem. 2002;277:30236–30243.
57. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the stability change of protein point mutations using neural networks. J Chem Inf Model. 2019;59:1508–1514.
58. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: An R package for the comparative analysis of protein structures. Bioinformatics. 2006;22:2695–2696.
59. Pires DE, Blundell TL, Ascher DB. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem. 2015;58:4066–4072.
60. Kaminskas LM, Pires DEV, Ascher DB. dendPoint: A web resource for dendrimer pharmacokinetics investigation and prediction. Sci Rep. 2019;9:15465.
61. Pires DEV, Ascher DB. mycoCSM: Using graph-based signatures to identify safe potent hits against mycobacteria. J Chem Inf Model. 2020;60:3450– 3456.
62. Rodrigues CH, Ascher DB, Pires DE. Kinact: A computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res. 2018;46:W127–W132.
63. Karmakar M, Rodrigues CHM, Holt KE, Dunstan SJ, Denholm J, Ascher DB. Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS One. 2019;14: e0217169.
64. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000;28:374.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science*. 2020;1–10. https://doi.org/10.1002/pro.3942

SUPPLEMENTARY MATERIAL

DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations

Carlos H.M. Rodrigues[1,2,*], Douglas E.V. Pires[1,2,3,#], David B. Ascher[1,2,4,#]

[1]Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Victoria, Australia

[2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Victoria, Australia

[3]School of Computing and Information Systems, University of Melbourne, Victoria, Australia

[4]Department of Biochemistry, University of Cambridge, Cambridge, UK

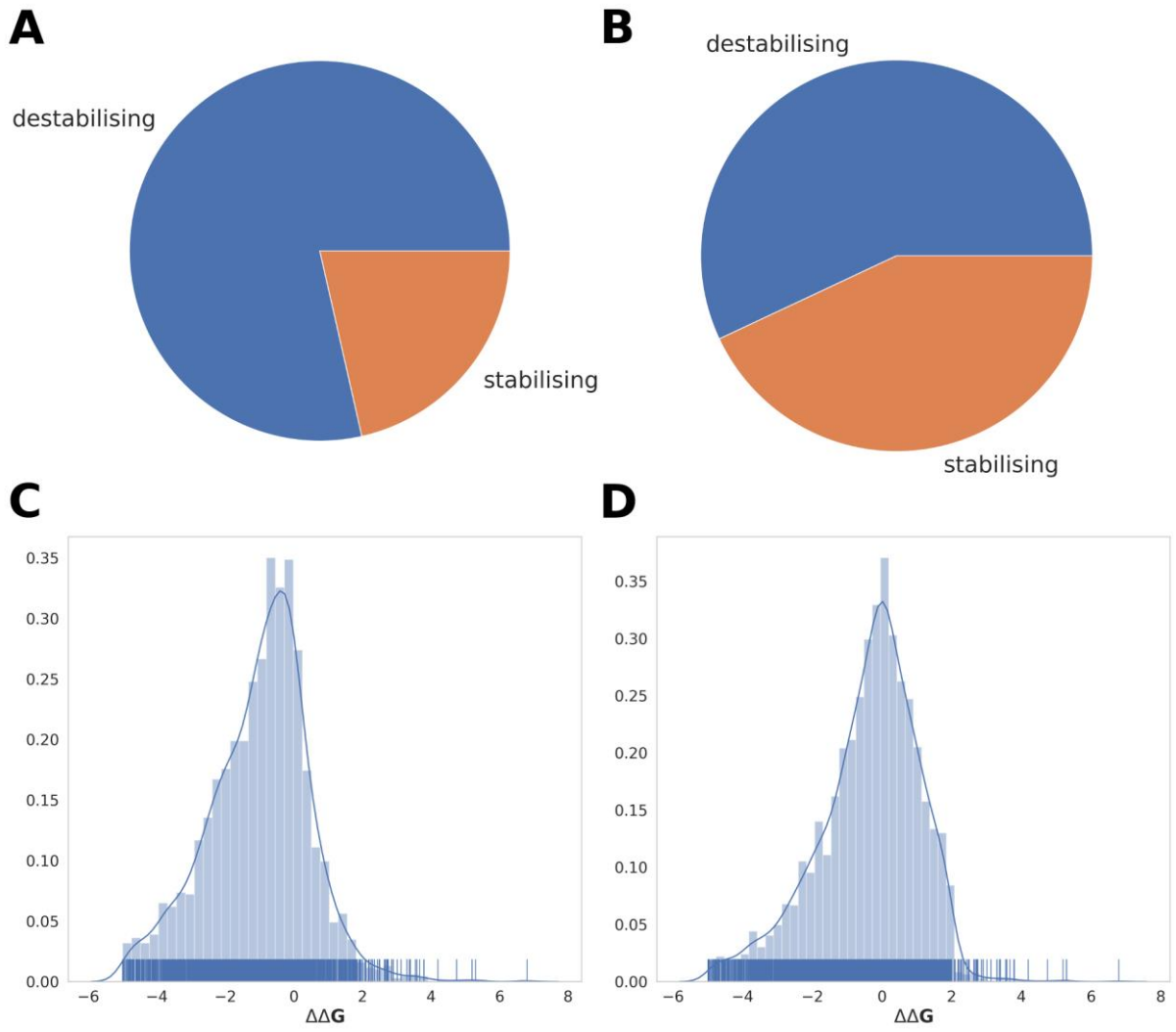


Figure S1 - Proportion of stabilising and destabilising mutations on S2648 and ΔΔG distribution. A) highlights the unbalanced nature of the original dataset with a much greater number of destabilising mutations over the stabilising ones. B) depicts the proportions for each one of the classes after including the reverse mutations. C) and D) shows the distribution of ΔΔG values for S2648 and our final dataset used in this study.

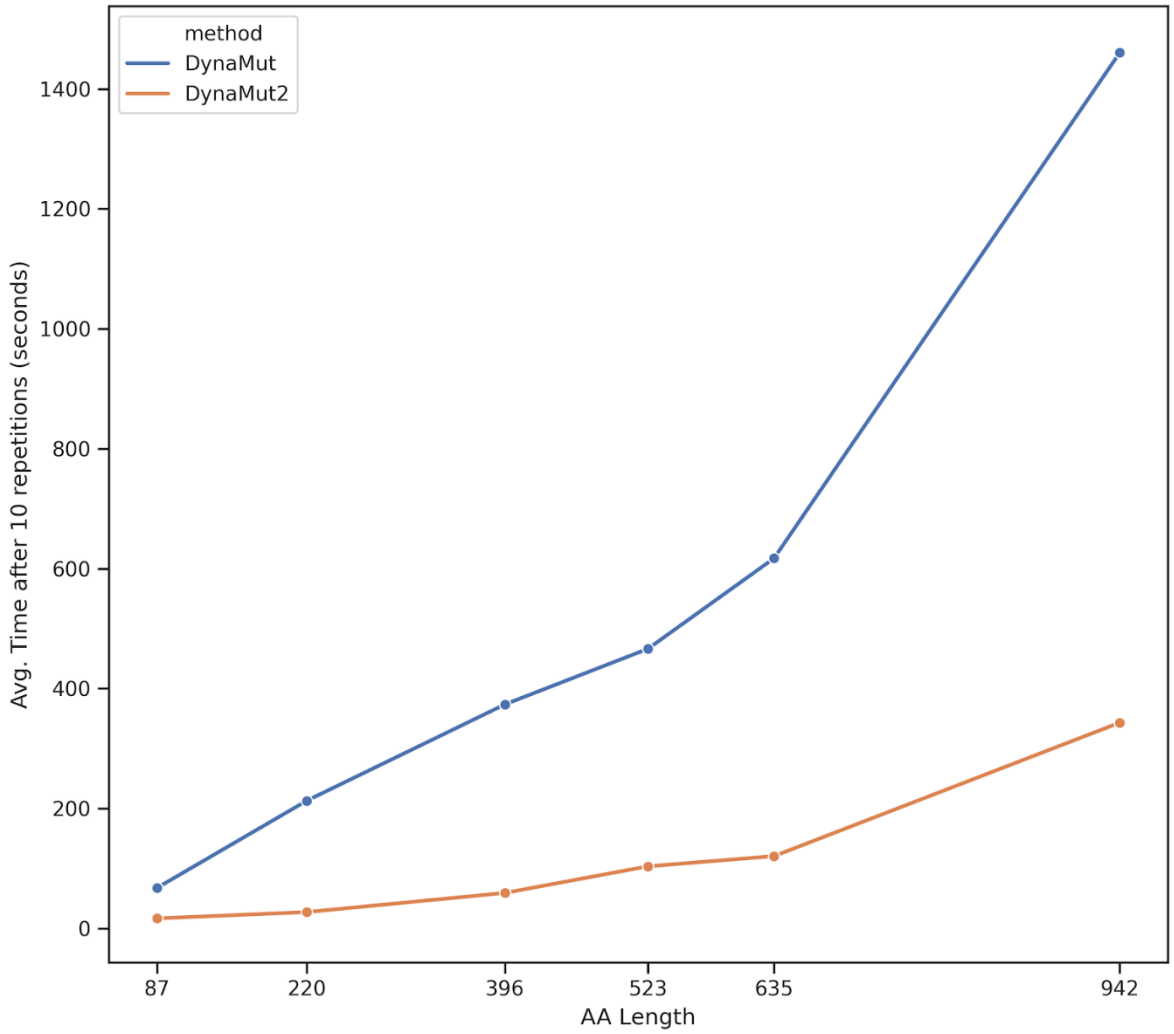


Figure S2 - Processing time of DynaMut and DynaMut2 on structures of different sizes. Here we show average values of processing time (in seconds) for Dynamut (blue) and DynaMut2 (orange) after 10 repetitions across 6 different proteins. DynaMut2 has a much lower processing time than its previous implementation on all cases, including a decrease of more than 6 times on larger structures. Details on structures and mutations are shown in Table S13.

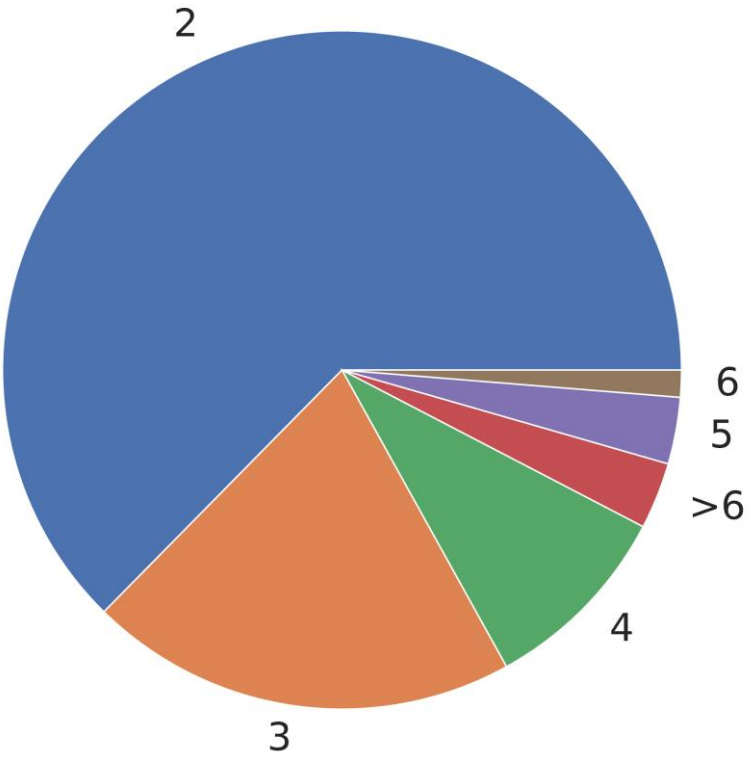


Figure S3 - Distribution of number of point mutations from the original dataset extracted from Protherm. More than 80% of the entries in the dataset comprises double and triple mutants.

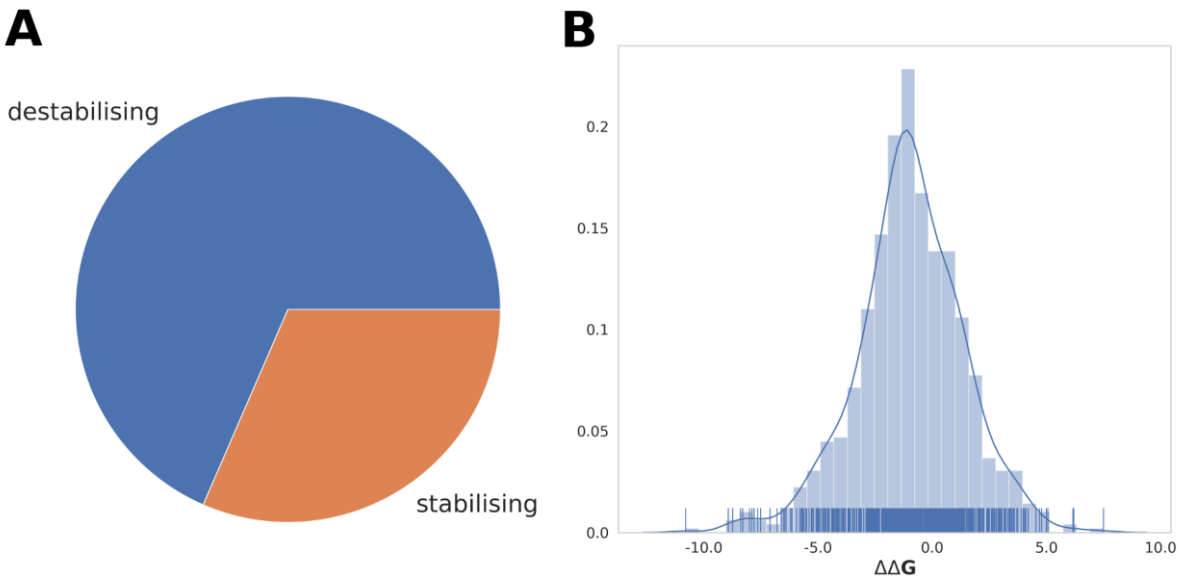


Figure S4 - Distribution of ΔΔG values for our dataset of multiple mutations.

Table S1 - Comparative performance on single-point mutation prediction on the test set S611 using rank correlation coefficients.

Methods	Overall		Stabilising		Destabilising	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
DynaMut2	**0.42**	**0.58**	**0.22**	**0.27**	**0.39**	**0.56**
DynaMut	0.27*	0.37+	0.11*	0.12+	0.36	0.52
DUET	0.25*	0.36+	0.01*	-0.01+	0.35	0.50
mCSM	0.23*	0.33+	0.01*	0.01+	0.33	0.47
SDM	0.23*	0.32+	0.09*	0.13+	0.24*	0.35+
ENCoM	-0.23*	-0.33+	-0.11*	-0.16+	-0.21*	-0.32+
Maestro	-0.15*	-0.23+	-0.0660*	-0.09+	-0.22*	-0.33+
I-Mutant	0.15*	0.22+	-0.07*	-0.11+	0.31*	0.44+
MUpro	0.07*	0.11+	-0.05*	-0.07+	0.13*	0.21+

* p-value < 0.05 by transforming tau-to-r followed by Fisher r-to-z transformation
+ p-value < 0.05 by transforming rho-to-r followed by Fisher r-to-z transformation

Table S2 - Comparative performance for single-point mutations on buried residues (RSA ≤ 30%) derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.41**	**1.42**
mCSM	0.24*	1.40
FoldX	0.21*	2.09
I-Mutant 3.0	0.14*	1.37
PoPMuSiC	0.19*	1.49
SDM	0.14*	1.55
Maestro	0.25*	1.41
CUPSAT	0.24*	1.87
Automute	0.26*	1.28

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S3 - Comparative performance for single-point mutations on exposed residues (RSA > 30%) derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.27**	**1.42**
mCSM	0.27	0.96
FoldX	0.16*	1.41
I-Mutant 3.0	0.15*	0.86
PoPMuSiC	0.16*	0.94
SDM	0.20	0.90
Maestro	0.37*	0.83
CUPSAT	0.16*	1.31
Automute	0.14*	1.43

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S4 - Comparative performance for single-point mutations on β-sheet structures according to CATH and derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.45**	**1.53**
mCSM	0.21*	1.66
FoldX	0.27*	2.23
I-Mutant 3.0	0.02*	1.78
PoPMuSiC	0.24*	1.66
SDM	-0.05*	1.89
Maestro	0.26*	1.66
CUPSAT	0.31*	2.05
Automute	0.08*	1.51

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S5 - Comparative performance for single-point mutations on α-helix and β-helix structures according to CATH and derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.37**	**0.99**
mCSM	0.25*	1.23
FoldX	0.25*	1.65
I-Mutant 3.0	0.24*	0.89
PoPMuSiC	0.27*	1.22
SDM	0.21*	1.34
Maestro	0.32	1.22
CUPSAT	0.23*	1.62
Automute	0.24*	1.39

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S6 - Comparative performance for single-point mutations on proteins with greater than 150 residues derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.43**	**1.47**
mCSM	0.18*	1.59
FoldX	0.21*	1.98
I-Mutant 3.0	0.11*	1.48
PoPMuSiC	0.07*	1.85
SDM	0.20*	1.61
Maestro	0.28*	1.54
CUPSAT	0.26*	2.06
Automute	0.24*	1.50

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S7 - Comparative performance for single-point mutations on proteins with less than 150 residues derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.40**	**1.02**
mCSM	0.27*	0.94
FoldX	0.24*	1.59
I-Mutant 3.0	0.27*	0.91
PoPMuSiC	0.37	0.90
SDM	0.23*	0.99
Maestro	0.37	0.89
CUPSAT	0.18*	1.32
Automute	0.22*	1.30

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S8 - Comparative performance for single-point mutations from large to small residues (in terms of volume) derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.37**	**1.66**
mCSM	0.15*	1.59
FoldX	0.21*	1.89
I-Mutant 3.0	0.12*	1.63
PoPMuSiC	0.07*	1.87
SDM	0.25*	2.06
Maestro	0.23*	1.67
CUPSAT	0.03*	2.39
Automute	0.21*	1.56

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S9 - Comparative performance for single-point mutations from small to large residues (in terms of volume) derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.47**	**1.40**
mCSM	0.09*	1.78
FoldX	0.10*	2.53
I-Mutant 3.0	-0.31*	1.52
PoPMuSiC	0.11*	1.74
SDM	-0.11*	1.72
Maestro	0.28*	1.82
CUPSAT	0.43	1.81
Automute	0.41	1.22

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S10 - Comparative performance for single-point mutations from and to residues with similar volumes derived from the O2567 non-redundant to training sets for each model.

Tool	Pearson	MAE
DynaMut2	**0.39**	**0.98**
mCSM	0.35	0.96
FoldX	0.24*	1.60
I-Mutant 3.0	0.33	0.89
PoPMuSiC	0.38	0.91
SDM	0.32	0.97
Maestro	0.40	0.88
CUPSAT	0.32	1.36
Automute	0.23*	1.31

** p-value < 0.05 compared with DynaMut2 using Fisher r-to-z transformation*

Table S11 - Pearson Correlation coefficient for performance over blind test S173 of experimental ΔTm.

Method	**1AQH**	**1H8V**	**1OSI**	**1XAS**	**2FJF**	**GK[a]**	**average**
DynaMut2	-0.33	0.25	0.23	0.35	0.40	0.45	0.23
DeepDDG	0.69	0.16	0.25	0.56	0.76	0.50	0.49
STRUM	0.24	0.09	0.30	0.34	0.62	0.39	0.33
SDM	-0.05	0.30	0.32	0.09	0.58	0.33	0.26
DUET	-0.43	0.33	0.36	0.41	0.39	0.37	0.24
mCSM	-0.55	0.22	0.32	0.33	0.34	0.34	0.17
I-Mutant	-0.29	0.03	0.32	0.30	0.25	0.30	0.15
MUpro	0.31	0.12	0.18	-0.47	0.30	0.33	0.13
DynaMut	-0.63	0.21	0.23	0.12	0.29	0.33	0.09

[a] *Guanylate kinase*

Table S12 - NMA force field options available on DynaMut2.

Name	**Description**
C-alpha (1)	Force field derived from fitting to the Amber94 all-atom potential
ANM (2)	Anisotropic Network Model uses a simplified spring force constant based on the pair-wise distance.
pfANM (3)	parameter-free Anisotropic Network Model is a variant from the ANM force field with interactions that fall off with the square of the distance.
REACH (4)	Realistic Extension Algorithm via Covariance Hessian is parameterized based on variance-covariance matrices obtained from MD simulations.
sdENM (5)	This force field employs residue specific spring force constants and it has been parameterized through a statistical analysis of 1500 NMR ensembles.

Table S13 - Summary of processing time for DynaMut and DynaMut2. Here we show average and standard deviation values after 10 repetitions for each mutation. Average values are shown with a confidence interval of 95%.

PDB	Mutation	Chain	DynaMut (avg.)	DynaMut (std.)	**DynaMut2 (avg.)**	**DynaMut2 (std.)**
1A43	G156A	A	67.50 ± 0.67*	1.08	**16.70 ± 1.17**	**1.89**
1AKY	V8I	A	212.90 ± 1.53*	2.47	**22.00 ± 0.41**	**0.67**
1AMQ	C270A	A	373.60 ± 2.67*	4.33	**43.25 ± 0.40**	**0.63**
1AON	T516V	A	466.30 ± 5.27*	8.50	**81.40 ± 2.86**	**4.62**
2ZT8	T526V	A	617.30 ± 2.50*	4.03	**112.10 ± 1.28**	**2.07**
6M71	Y884K	A	1460.50 ± 6.02*	9.71	**231.90 ± 3.42**	**5.53**

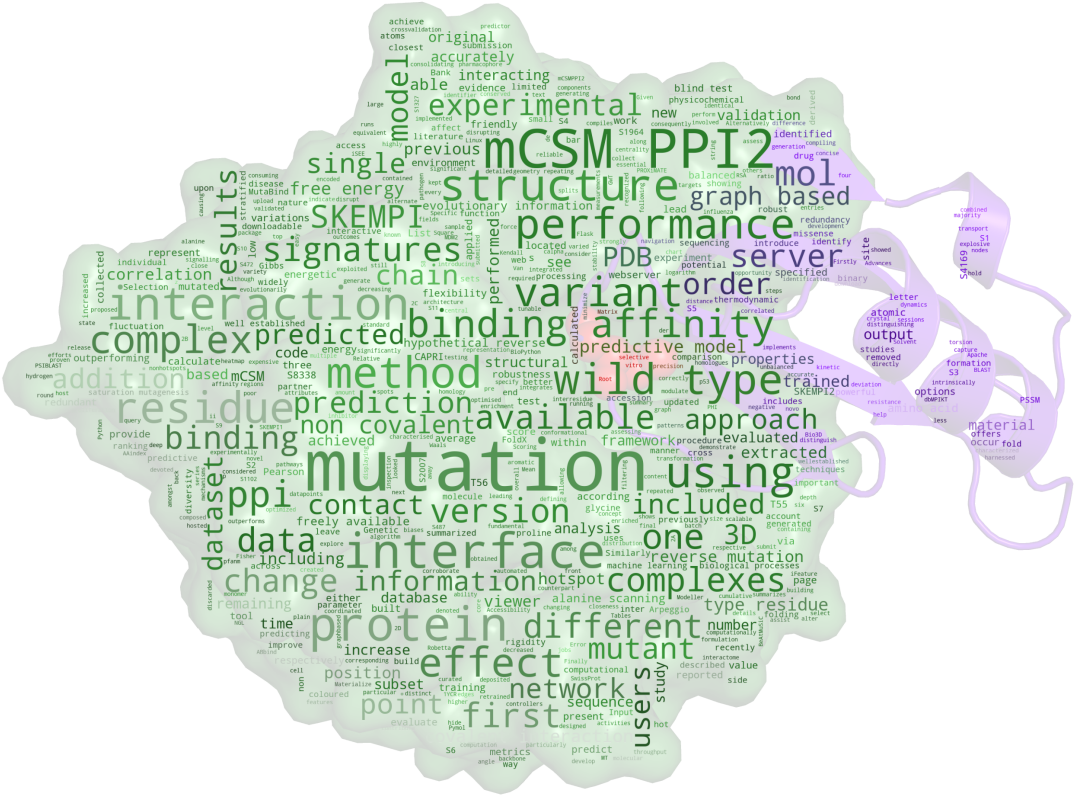
* p-value < 0.05 when evaluating average values with t-Test

REFERENCES

1. Hayward S, Kitao A, Go N (1995) Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. Proteins 23:177-186. PMID: 8592699 {Medline}
2. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505-515. PMID: 11159421 {Medline}
3. Yang L, Song G, Jernigan RL (2009) Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci U S A 106:12347-12352. PMID: 19617554 {Medline}
4. Moritsugu K, Smith JC (2007) Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. Biophys J 93:3460-3469. PMID: 17693469 {Medline}
5. Dehouck Y, Mikhailov AS (2013) Effective harmonic potentials: insights into the internal cooperativity and sequence-specificity of protein dynamics. PLoS Comput Biol 9:e1003209. PMID: 24009495 {Medline}

Chapter 4

Study of Effects of Single-Point Mutations on Protein-protein Interactions



W338–W344 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 22 May 2019*
doi: 10.1093/nar/gkz383

mCSM-PPI2: predicting the effects of mutations on protein–protein interactions

Carlos H.M. Rodrigues[1,2,3], Yoochan Myung[1,2,3], Douglas E.V. Pires [1,2,3,*] and David B. Ascher [1,2,3,4,*]

[1]Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Australia, [2]ACRF Facility for Innovative Cancer Drug Discovery, Bio21 Institute, University of Melbourne, Melbourne, Australia, [3]Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, Australia and [4]Department of Biochemistry, University of Cambridge, Cambridge, UK

Received February 11, 2019; Revised April 30, 2019; Editorial Decision April 30, 2019; Accepted May 20, 2019

ABSTRACT

Protein–protein Interactions are involved in most fundamental biological processes, with disease causing mutations enriched at their interfaces. Here we present mCSM-PPI2, a novel machine learning computational tool designed to more accurately predict the effects of missense mutations on protein–protein interaction binding affinity. mCSM-PPI2 uses graph-based structural signatures to model effects of variations on the inter-residue interaction network, evolutionary information, complex network metrics and energetic terms to generate an optimised predictor. We demonstrate that our method outperforms previous methods, ranking first among 26 others on CAPRI blind tests. mCSM-PPI2 is freely available as a user friendly webserver at http://biosig.unimelb.edu.au/mcsm_ppi2/.

INTRODUCTION

Most biological processes, including cell proliferation (1), signalling (2), host-pathogen interactions (3) and protein transport (4), are intrinsically coordinated through complex networks of protein–protein interactions. The diversity and size of the interactome offers a highly selective and tunable way to modulate protein activities and pathways (5). Genetic variations leading to changes in the binding affinity of these interactions can disrupt or directly affect the formation of interacting complexes and consequently lead to disease (6–16) and drug resistance (17–19).

Advances in next-generation sequencing techniques have created an explosive increase in the number of genetic variants available in the literature. However, experimental techniques to study these variants are still expensive and time consuming. mCSM (20) was one of the first scalable computational tools to accurately predict the effects of mutations on binding affinity. Previous methods were limited either in terms of their throughput (21,22) or in terms of their performance (23). Since then, significant efforts have been devoted to computationally study the effects of mutations on protein complexes (24,25) but their poor predictive performance on new variants, particularly mutations that lead to increased binding affinity of the complex, has limited their use. In addition, the increase in amount of experimental evidence of effects of variants on binding affinity offers the opportunity to develop new and more accurate methods.

Our previously described graph-based signatures concept has proven to be a powerful approach and has been widely applied to the study of protein structure, including how mutations alter protein stability (20,26), dynamics (27) and interactions with other molecules (20,28–34).

Here we introduce mCSM-PPI2, a webserver that integrates our well-established mCSM graph-based based signatures framework with evolutionary information, inter-residue non-covalent interaction networks analysis and energetic terms, in order to provide an optimized overall prediction performance.

MATERIALS AND METHODS

Data sets

The data used on this work was derived from the recently updated version of the SKEMPI database (35), which compiles experimental data on changes in thermodynamic and kinetic parameters on mutation for protein–protein complexes that have 3D structures deposited in the PDB. SKEMPI 2.0 (36) includes new mutations identified in the literature after its first release, including data available from three other databases: ABbind (37), PROXiMATE (38) and dbMPIKT (39). The average mutation effect was considered for variants reported in multiple experiments when these varied by less than 2.0 kcal/mol and discarded otherwise. After filtering for only single-point mutations with available

*To whom correspondence should be addressed. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au
Correspondence may also be addressed to Douglas E.V. Pires. Email: douglas.pires@unimelb.edu.au

experimental crystal structures of the wild-type, we were able to collect 4169 (S4169) variants in 319 different complexes. All protein structures were collected from the Protein Data Bank and a series of pre-processing steps were performed to account for the diversity of structures (see Supplementary material).

The binding affinity of protein–protein complexes were used to calculate the binding Gibbs free energy (ΔG):

$$\Delta G = RT \ln (K_D)$$

where $R = 1.985 \times 10^{-3}$ kcal $K^{-1}mol^{-1}$ is the ideal gas constant, T is the temperature (in K) and K_D is the equilibrium dissociation constant of the protein–protein complex (in molar). The change in binding affinity upon mutation was calculated as follows:

$$\Delta\Delta G_{wt-mt} = \Delta G_{wild-type} - \Delta G_{mutant}$$

Since the Gibbs free energy formulation is a thermodynamic state function a change in binding affinity of a mutation from a wild-type protein to its mutant ($\Delta\Delta G_{WT \rightarrow MT}$) should be equivalent to the negative change binding free energy of the hypothetical reverse mutation, from the mutant to the wild-type protein ($-\Delta\Delta G_{MT \rightarrow WT}$) (40). Given the unbalanced nature of the original dataset collected from SKEMPI 2.0, 901 variants increased ($\Delta\Delta G_{wt\text{-}mt} \geq 0$) and 3268 decreased ($\Delta\Delta G_{wt\text{-}mt} < 0$) binding affinity, and in order to build a more robust and balanced predictive method, we also included the hypothetical reverse mutations. Therefore, the final dataset for building mCSM-PPI2 predictive model includes 8338 single-point mutations (S8338), which represents an increase of up to three-fold in datapoints in comparison with previous methods that used data from the first version of SKEMPI with 2007 (S2007) (20,23), 1964 (S1964) (25), 1102 (S1102) (24) and 1327 (S1327) mutations (41).

A subset of 487 mutations in 56 complexes (S487) contained within S4169 and not in S2007 were recently curated (24) and here we used as evidence to evaluate the performance of mCSM-PPI2. A summary of different subsets derived from SKEMPI is shown in Supplementary Table S1.

The datasets used in this work are freely available for download at http://biosig.unimelb.edu.au/mcsm_ppi2/datasets.

Graph-based structural signatures

mCSM-PPI2 uses as one of its core components our well established graph-based structural signatures (mCSM) to represent the environment of the wild-type residue. This approach models both the geometry and physicochemical properties of the interactions and architecture of wild-type structure and has been widely applied to the study of small molecule and protein structure (20,26–34,42). Our signatures represent atoms as nodes and their interactions as edges, with their physicochemical properties encoded based upon the amino acid residue properties, denoted by a pharmacophore. From this representation of the residue environment, distance patterns between atoms characterized by their properties are summarized in concise signatures as cumulative distribution functions.

Modelling effects of mutation

Single-point mutations can affect protein–protein interactions via different molecular mechanisms, including changing folding free energy of interacting partners or disrupting non-covalent interactions essential for complex formation (6,43). In mCSM-PPI2, we have included six new distinct types of features that were not used in our first method (Supplementary Table S2). These were combined with our well-established graph-based signatures as evidence for training a machine learning algorithm (see Supplementary material) to better explore the effects of mutations in protein–protein binding affinity (Figure 1).

Wild-type residue environment. Based on 3D structures collected from the Protein Data Bank (44), we were able to calculate Relative Solvent Accessibility (RSA), torsion angle PHI and residue depth for the wild-type residue using BioPython (45) version 1.7. We also extracted information on the amino acid content in the sequence of the chain in which the mutation occurs using iFeature (46).

Nature of wild-type and mutant residues. The conformational flexibility of glycine side chains and the rigidity of proline side chains are important for defining the backbone flexibility. Mutations from and to these two amino acids can lead to large structural effects. For our model we included binary terms to capture if the mutation was from or to a glycine or proline.

Evolutionary information. Binding regions are known to be evolutionarily conserved, which has been exploited in a variety of studies to identify potential protein interaction interfaces. For mCSM-PPI2 we also harnessed this information by using the Position Specific Scoring Matrix (PSSM) scores. PSSM was calculated through PSI-BLAST of BLAST 2.2.3 using the non-redundant Swiss-Prot database of protein sequences and the sequence of the chain in which the mutation occurs as the query parameter.

Non-covalent interaction network analysis. We performed analysis of the non-covalent interactions for the wild-type residue and for the closest interface using the contacts calculated by Arpeggio (47). Here, we extracted two types of information: the difference between the number of contacts of wild-type and mutant residue for covalent, Van der Waals', aromatic and hydrogen bond contacts, and complex network metrics for the contact graph of the closest interface of interaction, from which we extracted centrality metrics, including closeness and central points (48). In this work, we consider a residue to be at an interaction interface if it is located at most 5 Å away from the interacting partner, following previous studies. In addition, we included three protein contact potentials scores from the AAindex database (49) (Supplementary Table S3).

Energetic terms. Interaction energy information between the two interacting chains were extracted from FoldX (22). In addition, we included the predicted folding free energy change upon mutation.

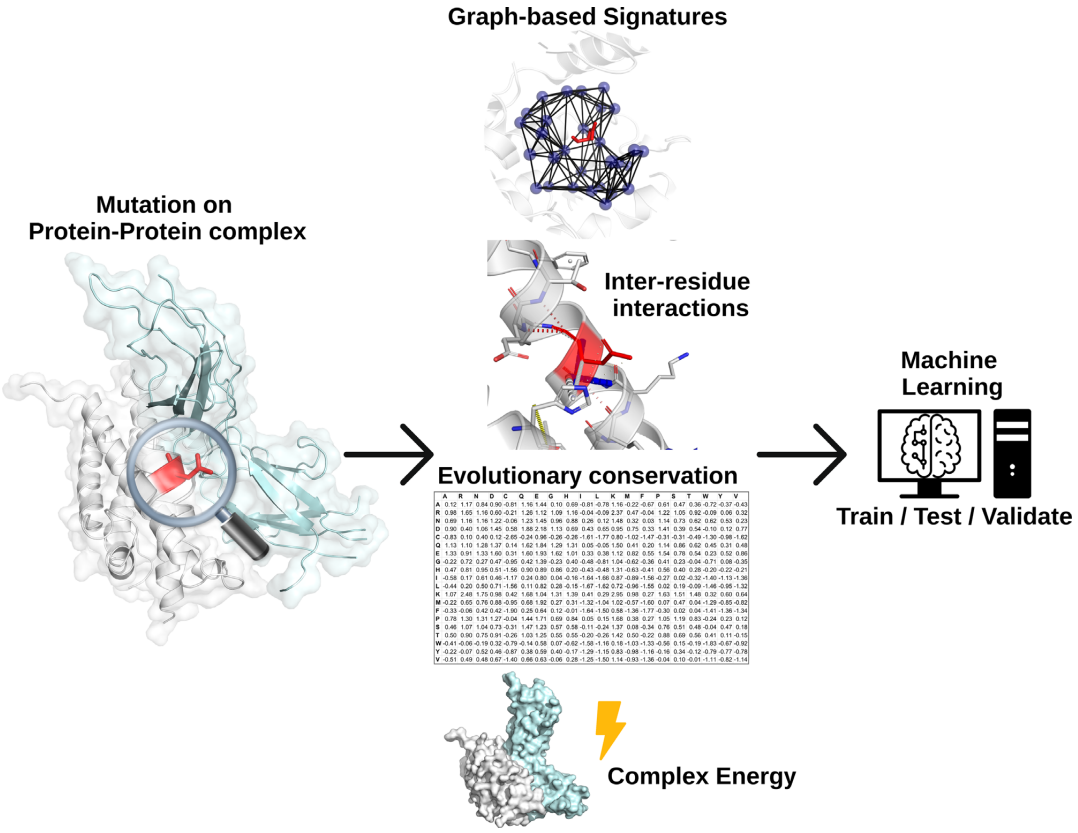


Figure 1. mCSM-PPI2 methodology workflow. The method relies on graph-based signatures, which model distance patterns and encode geometrical and physico-chemical properties on wild-type residue environment. Network analysis based on non-covalent interactions of wild-type residue and interacting interface along with evolutionary information and energy terms are also used. All features are used as evidence to train, test and validate machine learning algorithms.

Atomic fluctuation. We used the Bio3D R package (50) to calculate the atomic fluctuations of the structure of the monomer where the mutation occur using calpha and pfanm force fields to account for effects on protein flexibility/rigidity.

WEBSERVER

We have implemented mCSM-PPI2 as a user-friendly and freely available webserver (http://biosig.unimelb.edu.au/mcsm_ppi2/). The server front end was built using Materialize framework version 1.0.0, while the back-end was built in Python via the Flask framework (Version 1.0.2). It is hosted on a Linux server running Apache.

Input

mCSM-PPI2 can be used in two different ways: to either assess the effects of mutations specified by the user input or to predict the effects of mutations at the protein–protein interface in an automated manner. For user-specified variations two options are available (Supplementary Figure S1). The 'Single Mutation' option requires one to provide a PDB file or PDB accession code of the structure of the protein complex, the point mutation specified as a string containing the wild-type residue one-letter code, its corresponding residue number and the mutant residue one-letter code. The 'Mutation List' option allows users to upload a list of mutations in a plain text file for batch processing. For both options, users are also required to specify the chain identifier in which the wild-type residues are located.

Alternatively, for assessing effects of mutations at protein–protein interfaces the server requires the user to provide a PDB file or PDB accession code and select one of two options: alanine scanning (all interface residues are mutated to an Alanine) or saturation mutagenesis (all interface residues are mutated to every other amino acid) (Supplementary Figure S2).

In order to assist users to submit their jobs for predictions, sample submission entries are available in both submission pages and a help page is also available via the top navigation bar.

Output

For the Single Mutation option (Supplementary Figure S3), mCSM-PPI2 outputs the predicted change in binding affinity (in kcal/mol) along with an interactive 3D viewer, built using NGL viewer (51), showing non-covalent interactions, generated with Arpeggio, at the mutated position. A set of controllers are available for users to hide and show the different types of interactions and to alternate between wild-

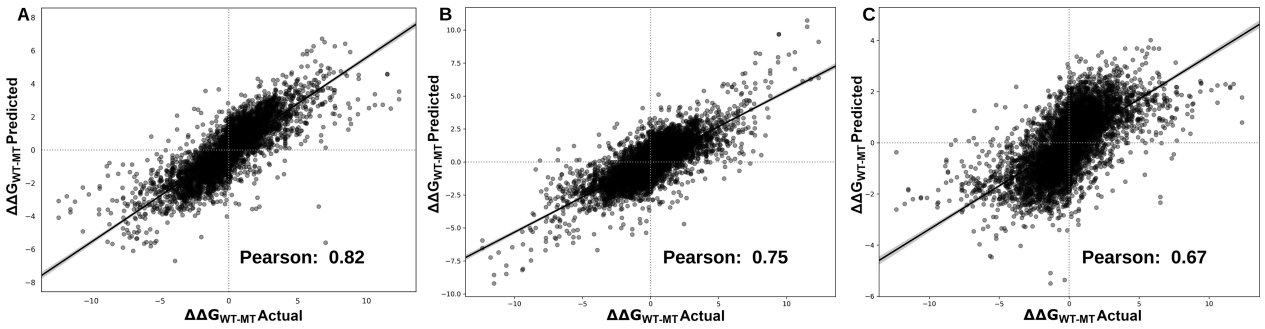


Figure 2. Performance evaluation on cross-validation. mCSM-PPI2 was able to achieve a Pearson's correlation of 0.82 and RMSE 1.18 kcal/mol when trained on the S8338 dataset applying 10-fold cross-validation 10 times (**A**). In low-redundancy sets, mCSM-PPI2 was able to achieve a correlation of 0.75 and 0.67 on leave-one-complex-out (**B**) and leave-one-binding-site-out (**C**), respectively.

type and mutant structures. In addition, a 2D viewer displaying non-covalent interactions of wild-type and mutant structures is also shown. Pymol sessions are available for download. For the Mutation List option (Supplementary Figure S4), the results are summarized in a downloadable table from which users can access details for each single variant.

For the Alanine Scanning option on the interface analysis, the server first presents a table with all the interfaces identified on the submitted structure, and it also allows for inspection of the individual interfaces. On the results page of each interface the server shows a downloadable table with the prediction outcomes for each mutation, a bar chart that summarizes the predicted changes in binding affinity (Supplementary Figure S5) and an interactive 3D viewer in which the residues are coloured according to the predicted value (Supplementary Figure S6). Similarly, for the Saturation Mutagenesis option, mCSM-PPI2 outputs a table with all the interfaces identified and allows the users to access detailed information on each interface. For each interface, the server outputs a table compiling the results for all variants (Supplementary Figure S7), a heatmap of all interface residues and their respective mutations (Supplementary Figure S8), and a 3D viewer in which the residues are coloured according to the average prediction for each particular position (Supplementary Figure S9).

VALIDATION

Performance on cross-validation

In order to build a more robust and reliable predictive model we performed four types of validation. Firstly, we performed 10-times stratified 10-fold cross-validation, using 90% of our original dataset (S8338) for training and the remaining as a blind test. Selection of the blind test was repeated 10 times in a stratified manner, with the model retrained on the remaining data, in order to test the robustness of the model (see Supplementary material). For this approach the hypothetical reverse mutations were kept in either training or test sets during the splits according to its counterpart original mutation. Our method was able to achieve an average Pearson correlation (ρ) of 0.82 with a standard deviation (σ) of 0.06 across the 10 runs (Figure 2A) showing a more balanced prediction when distinguishing between mutations that increase binding affinity from decreasing ones than other methods (Supplementary Table S4). We also evaluate the performance of mCSM-PPI2 when trained only on the original subset of mutations from SKEMPI2 (S4169) using the same procedure and obtained a correlation of 0.76 and RMSE of 1.19 kcal/mol. These results corroborate the use of reverse mutations in order to improve performance and robustness of our predictive model. Performance comparison between mCSM-PPI2 and other methods on different versions of SKEMPI and performance of individual types of attributes are shown in Supplementary Tables S5 and S6, respectively.

We further evaluated the performance of our approach on two low-redundancy sets; low redundant at the (i) complex and (ii) interface level. The complex low redundancy test was performed using leave-one-complex out cross-validation, in which we trained our model on 318 complexes of our dataset and evaluate the performance on the one remaining complex. After repeating this procedure for each complex we achieved $\rho = 0.75$ (Figure 2B) and Root Mean Square Error (RMSE) of 1.30 kcal/mol, outperforming MutaBind (25) ($\rho = 0.68$ and RMSE = 1.57 kcal/mol).

Similarly, we applied leave-one binding site out using the 'hold-out' information extracted from SKEMPI2. Here, we removed all mutations located in identical binding sites for testing and trained on the remaining data. mCSM-PPI2 achieved $\rho = 0.67$ (RMSE = 1.39 kcal/mol) (Figure 2C), which was significantly higher (p-value < 0.0001 by Fisher *r*-to-*z* transformation) than the results reported for MutaBind when trained using only mutations from SKEMPI1 ($\rho = 0.57$ and RMSE = 1.57 kcal/mol).

In addition, we evaluated the performance of our approach on a subset of 472 mutations (S472) not present within the first version of SKEMPI but included in SKEMPI2. For this experiment, we trained a predictive model using all variants from the first version of SKEMPI (S1964). Our method achieved a correlation of 0.63 (RMSE = 1.11 kcal/mol).

Validation on CAPRI

mCSM-PPI2 was further validated against the CAPRI (52) round 26, which is composed of 1862 experimentally characterised mutations in two *de novo* influenza inhibitor targets (T55 and T56: 1007 mutations at 53 different positions

Table 1. Comparative performance of mCSM-PPI2 on CAPRI and the blind test set for the complex MDM2-P53

	CAPRI (T55)		CAPRI (T56)		MDM2-P53	
Method	Kendall	RMSE (kcal/mol)	Kendall	RMSE (kcal/mol)	ρ	RMSE (kcal/mol)
mCSM-PPI2	**0.42**	**2.55**	**0.32**	**4.06**	**0.40**	**0.36**
mCSM (20)	0.16**	3.71	0.13**	4.15	0.23	0.83
MutaBind (25)	0.41	2.58	0.30	4.27	*NA*	*NA*
iSEE (24)	*NA*	*NA*	*NA*	*NA*	0.24	0.81
BeAtMuSiC (23)	0.28**	3.04	0.30	4.06	−0.23*	0.91
FoldX (22)	0.12**	3.94 0.90	0.16**	4.33	−0.14*	
MMPBSA (21)	0.19**	5.40	0.08**	28.04	*NA*	*NA*

*p-value < 0.05 by Fisher r-to-z transformation test compared to mCSM-PPI2
**p-value < 0.05 by transforming tau-to-r followed by Fisher r-to-z transformation. *NA*: Data not available.

in T55 and 855 mutations at 45 different positions in T56). The *in vitro* experimental measurements used the enrichment values generated from deep sequencing and were calculated based on the binary logarithm of the ratio of number of times the variant sequence was observed after and before the selection for binding. Although the 3D structures for these two complexes were not available, structures of close homologues have been described (53,54) and were used for generating homology model structures by introducing point mutations using Modeller (55) (see Supplementary Materials). mCSM-PPI2 was able to achieve a Kendall's score of up to 0.42 and 0.32 for mutations in T55 and T56, respectively, ranking first amongst 26 other methods (Supplementary Figure S10 and Table 1).

Blind test

The performance of mCSM-PPI2 was further evaluated on a small set of 26 variants at the interface of interaction of the MDM2-p53 complex (PDB 1YCR) (24). Our method achieved a Pearson's Correlation of 0.40 and an RMSE = 0.36 kcal/mol outperforming mCSM, iSEE, FoldX, BeAtMuSiC (23) (Table 1).

Finally, we looked at the ability of mCSM-PPI2 to accurately identify PPI hotspots, residues that contribute to the majority of the binding free energy of the interaction and have been recognized as important sites for drug development (5). Here we evaluated the performance of mCSM-PPI2 across a previously proposed set of 378 alanine-scanning experimental mutations within 19 different protein–protein complexes (56,57) (Supplementary Table S7). In order to minimize biases, for this experiment we removed 232 variants from S8338 which were redundant with our set of 378 alanine scanning mutations. Our predictive model was was able to accurately distinguish hot and not-hot spots (95% of hotspots and 92% of non-hotspots were correctly predicted) outperforming the results reported for Robetta (precision of 79% and 68% when predicting hotspots and non-hotspots, respectively). The predicted changes in binding energy showed that mCSM-PPI2 predictions correlated strongly with the experimental data (Pearson's Correlation of 0.95 and RMSE of 0.25 kcal/mol; Supplementary Figure S11). These results indicate that mCSM-PPI2 could also be a powerful tool for hotspot identification.

CONCLUSION

Here, we introduce mCSM-PPI2, a web server that implements an integrated computation approach for predicting effects of missense mutations in protein–protein affinity. By consolidating our graph-based signatures framework with evolutionary information, inter-atomic contacts and energy terms our updated method has shown to perform better than its previous version and other methods. In addition, the use of hypothetical reverse mutations has shown to improve the robustness of our predictive model allowing for a more balanced prediction. mCSM-PPI2 is freely available as a user-friendly and easy to use web server at http://biosig.unimelb.edu.au/mcsm_ppi2/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Australian Government Research Training Program Scholarship [to C.H.M.R and Y.M.]; Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; Victorian Life Sciences Computation Initiative (VLSCI), an initiative of the Victorian Government, Australia, on its Facility hosted at the University of Melbourne [UOM0017]; Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P.]; Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.]; Supported in part by the Victorian Government's OIS Program. Funding for open access charge: MRC.
Conflict of interest statement. None declared.

REFERENCES

1 Gao,J., Li,W.X., Feng,S.Q., Yuan,Y.S., Wan,D.F., Han,W. and Yu,Y. (2008) A protein–protein interaction network of transcription factors acting during liver cell proliferation. *Genomics*, **91**, 347–355.

2 Chuderland,D. and Seger,R. (2005) Protein-protein interactions in the regulation of the extracellular signal-regulated kinase. *Mol. Biotechnol.*, **29**, 57–74.
3 Nicod,C., Banaei-Esfahani,A. and Collins,B.C. (2017) Elucidation of host-pathogen protein–protein interactions to uncover mechanisms of host cell rewiring. *Curr. Opin. Microbiol.*, **39**, 7–15.
4 Paumi,C.M., Menendez,J., Arnoldo,A., Engels,K., Iyer,K.R., Thaminy,S., Georgiev,O., Barral,Y., Michaelis,S. and Stagljar,I. (2007) Mapping protein–protein interactions for the yeast ABC transporter Ycf1p by integrated split-ubiquitin membrane yeast two-hybrid analysis. *Mol. Cell*, **26**, 15–25.
5 Jubb,H., Blundell,T.L. and Ascher,D.B. (2015) Flexibility and small pockets at protein–protein interfaces: New insights into druggability. *Prog. Biophys. Mol. Biol.*, **119**, 2–9.
6 Gao,M., Zhou,H and Skolnick,J. (2015) Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure*, **23**, 1362–1369.
7 David,A., Razali,R., Wass,M.N. and Sternberg,M.J. (2012) Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.*, **33**, 359–363.
8 Engin,H.B., Kreisberg,J.F. and Carter,H. (2016) Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PLoS One*, **11**, e0152929.
9 Jubb,H.C., Pandurangan,A.P., Turner,M.A., Ochoa-Montano,B., Blundell,T.L. and Ascher,D.B. (2017) Mutations at protein–protein interfaces: small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Biol.*, **128**, 3–13.
10 Ascher,D.B., Spiga,O., Sekelska,M., Pires,D.E.V., Bernini,A., Tiezzi,M., Kralovicova,J., Borovska,I., Soltysova,A., Olsson,B. *et al.* (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.*, **27**, 888–902.
11 Hnizda,A., Fabry,M., Moriyama,T., Pachl,P., Kugler,M., Brinsa,V., Ascher,D.B., Carroll,W.L., Novak,P., Zaliova,M. *et al.* (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. *Leukemia*, **32**, 1393–1403.
12 Andrews,K.A., Ascher,D.B., Pires,D.E.V., Barnes,D.R., Vialard,L., Casey,R.T., Bradshaw,N., Adlard,J., Aylwin,S., Brennan,P. *et al.* (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J. Med. Genet.*, **55**, 384–394.
13 Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B., Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom. Med.*, **2**, 7.
14 Nemethova,M., Radvanszky,J., Kadasi,L., Ascher,D.B., Pires,D.E., Blundell,T.L., Porfirio,B., Mannoni,A., Santucci,A., Milucci,L. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur. J. Hum. Genet.*, **24**, 66–72.
15 Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.
16 Blaszczyk,M., Harmer,N.J., Chirgadze,D.Y., Ascher,D.B. and Blundell,T.L. (2015) Achieving high signal-to-noise in cell regulatory systems: Spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. *Prog. Biophys. Mol. Biol.*, **118**, 103–111.
17 Ascher,D.B., Wielens,J., Nero,T.L., Doughty,L., Morton,C.J. and Parker,M.W. (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci. Rep.*, **4**, 4765.
18 Portelli,S., Phelan,J.E., Ascher,D.B., Clark,T.G. and Furnham,N. (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Sci. Rep.*, **8**, 15356.
19 Vedithi,S.C., Malhotra,S., Das,M., Daniel,S., Kishore,N., George,A., Arumugam,S., Rajan,L., Ebenezer,M., Ascher,D.B. *et al.* (2018) Structural implications of mutations conferring rifampin resistance in mycobacterium leprae. *Sci. Rep.*, **8**, 5016.
20 Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
21 Kollman,P.A., Massova,I., Reyes,C., Kuhn,B., Huo,S., Chong,L., Lee,M., Lee,T., Duan,Y., Wang,W. *et al.* (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, **33**, 889–897.
22 Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
23 Dehouck,Y., Kwasigroch,J.M., Rooman,M. and Gilis,D. (2013) BeAtMuSiC: Prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res.*, **41**, W333–W339.
24 Geng,C., Vangone,A., Folkers,G.E., Xue,L.C. and Bonvin,A. (2019) iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins*, **87**, 110–119.
25 Li,M., Simonetti,F.L., Goncearenco,A. and Panchenko,A.R. (2016) MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic Acids Res.*, **44**, W494–W501.
26 Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–W319.
27 Rodrigues,C.H., Pires,D.E. and Ascher,D.B. (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.*, **46**, W350–W355.
28 Rodrigues,C.H., Ascher,D.B. and Pires,D.E. (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. *Nucleic Acids Res.*, **46**, W127–W132.
29 Pires,D.E.V. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.
30 Pires,D.E., Chen,J., Blundell,T.L. and Ascher,D.B. (2016) In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.*, **6**, 19848.
31 Pires,D.E., Blundell,T.L. and Ascher,D.B. (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.*, **6**, 29575.
32 Pires,D.E. and Ascher,D.B. (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res.*, **44**, W557–W561.
33 Pires,D.E. and Ascher,D.B. (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.*, **44**, W469–W473.
34 Pires,D.E., Blundell,T.L. and Ascher,D.B. (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res.*, **43**, D387–D391.
35 Moal,I.H. and Fernandez-Recio,J. (2012) SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.
36 Jankauskaite,J., Jimenez-Garcia,B., Dapkunas,J., Fernandez-Recio,J. and Moal,I.H. (2019) SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, **35**, 462–469.
37 Sirin,S., Apgar,J.R., Bennett,E.M. and Keating,A.E. (2016) AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Sci.*, **25**, 393–409.
38 Jemimah,S., Yugandhar,K. and Michael Gromiha,M. (2017) PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, **33**, 2787–2788.
39 Liu,Q., Chen,P., Wang,B., Zhang,J. and Li,J. (2018) dbMPIKT: a database of kinetic and thermodynamic mutant protein interactions. *BMC Bioinformatics*, **19**, 455.
40 Thiltgen,G. and Goldstein,R.A. (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One*, **7**, e46084.
41 Petukh,M., Dai,L. and Alexov,E. (2016) SAAMBE: Webserver to predict the charge of binding free energy caused by amino acids mutations. *Int. J. Mol. Sci.*, **17**, 547.
42 Pires,D.E., Blundell,T.L. and Ascher,D.B. (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem.*, **58**, 4066–4072.

43 Sahni,N., Yi,S., Taipale,M., Fuxman Bass,J.I., Coulombe-Huntington,J., Yang,F., Peng,J., Weile,J., Karras,G.I., Wang,Y. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.
44 Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
45 Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
46 Chen,Z., Zhao,P., Li,F., Leier,A., Marquez-Lago,T.T., Wang,Y., Webb,G.I., Smith,A.I., Daly,R.J., Chou,K.C. *et al.* (2018) iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34**, 2499–2502.
47 Jubb,H.C., Higueruelo,A.P., Ochoa-Montano,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.
48 Li,G., Semerci,M., Yener,B. and Zaki,M.J. (2012) Effective graph classification based on topological and label attributes. *Stat. Anal. Data Mining: ASA Data Sci. J.*, **5**, 265–283.
49 Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.
50 Grant,B.J., Rodrigues,A.P., ElSawy,K.M., McCammon,J.A. and Caves,L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.
51 Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlic,A. and Rose,P.W. (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
52 Janin,J., Henrick,K., Moult,J., Eyck,L.T., Sternberg,M.J., Vajda,S., Vakser,I., Wodak,S.J. and Critical Assessment of, P.I. (2003) CAPRI: a critical assessment of predicted interactions. *Proteins*, **52**, 2–9.
53 Fleishman,S.J., Whitehead,T.A., Ekiert,D.C., Dreyfus,C., Corn,J.E., Strauch,E.M., Wilson,I.A. and Baker,D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.
54 Whitehead,T.A., Chevalier,A., Song,Y., Dreyfus,C., Fleishman,S.J., De Mattos,C., Myers,C.A., Kamisetty,H., Blair,P., Wilson,I.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.*, **30**, 543–548.
55 Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
56 Ascher,D.B., Jubb,H.C., Pires,D.E., Ochi,T., Higueruelo,A. and Blundell,T.L. (2015) Protein-protein interactions: structures and druggability. In: Scapin,G, Patel,D and Arnold,E (eds). *Multifaceted Roles of Crystallography in Modern Drug Discovery*. Springer, Netherlands, pp. 141–163.
57 Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 14116–14121.

SUPPLEMENTARY MATERIAL

mCSM-PPI2: predicting the effects of mutations on protein-protein interactions

Carlos H.M. Rodrigues[1,2,3], Yoochan Myung[1,2,3], Douglas E.V. Pires[1,2,3,*], David B. Ascher[1,2,3,4,*]

[1]Department of Biochemistry and Molecular Biology, University of Melbourne;
[2]ACRF Facility for Innovative Cancer Drug Discovery, Bio21 Institute, University of Melbourne;
[3]Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute
[4]Department of Biochemistry, University of Cambridge;

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au or da382@cam.ac.uk. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

Machine Learning

mCSM-PPI2 was built using the open source module Scikit-Learn version 0.20.1 (1). Scikit-Learn provides a collection of simple and powerful functions, tools and algorithms to perform machine learning and data analysis in Python. Our final model and the results presented in this work were generated based on the ExtraTrees algorithm, a tree-based ensemble method for supervised classification and regression tasks (2). This algorithm operates in a similar manner as the Random Forest algorithm with the difference that decisions boundaries used when splitting the nodes for building the trees are chosen randomly. When building our predictive models we used the default parameters defined in Scikit-Learn except for the number of instances, which was set to 300.

Stratified Cross-Validation

Results for 10-fold cross-validation performance of mCSM-PPI2 were generated after 10 times repetition of the traditional cross-validation in a stratified manner (pseudo-code below). The dataset used for training was divided in 10 subsets and for each repetition one subset was used for testing and the remaining data was used for training in which 10-fold cross-validation was carried out. Therefore, every subset was used as test set only one time. This procedure was used in order to reduce redundancy over the repetitions.

Pseudo-code for stratified cross-validation:

```
N = 10
K = 10
subsets = split(original_dataset, N)
for each subset:
    remaining_dataset = original_dataset - subset
    performance_on_crossvalidation(original_dataset - subset, K)
    model = train_predictive_model(remaining_dataset)
    model.evaluate(subset)
```

Homology Model Structures for CAPRI Dataset

3D structures for the *de novo* designed influenza inhibitors (HB36.4 and HB80.3) in complex with hemagglutinin (HA), specified on the 26th round of CAPRI as targets T55 and T56, were obtained via homology modelling using Modeller version 9.21 (3) and close homologues previously defined (4,5). The 3D structure for Target 55 were obtained by introducing a single point mutation (N64K) on the structure of HB36.3-HA (PDB: 3R2X). For Target 56, we used the crystal structure of the complex HB80.4-HA (PDB: 4EEF) as the template and introduced 5 point mutations (K12G, I17L, I21L, K35A, K42S).

Pre-Processing of PDB Structures

Due to limitations and eccentricities of the PDB format, we performed a series of steps to ensure that structures could be processed by external software. These are described as follows:

- If an accession code for a protein structure on the Protein Data Bank is provided on the input page, mCSM-PPI2 will try to download the first author assigned biological assembly defined in the headers of the official structure. If no code is found (*e.g.*, on NMR structures), mCSM-PPI2 will use the official structure file available on the Protein Data Bank.
- Where multiple models were present in a structure (*e.g.*, on NMR structures), only the first one was maintained, ensuring only a static state was compared when conducting large scale analysis.
- Water molecules, ligands and non-standard compounds were removed.
- For residues presenting multiple occupancy, only the most prevalent one was selected.
- Given that some of the tools used by mCSM-PPI2 have difficulties dealing with structures with insertion codes, temporary files with residues renumbered sequentially (starting from 0) before running our calculations.
- Missing atoms and residues were not modelled.

TABLES

Table S1 – Summary of datasets of experimental information on point mutations and their effects on binding affinity.

Dataset	Description	Reference
S4169	Variants extracted from Skempi2	(16)
S8338	S4169 plus all reverse mutations	(16)
S2007	2007 variants extracted from Skempi1	(10,17,18)
S1964	1964 variants extracted from Skempi1	(17,19)
S1327	1327 variants extracted from Skempi1	(17,20)
S1102	1102 variants extracted from Skempi1	(17,21)
S472	472 variants contained in S4169 and not in S2007	(16,21)
S378	Alanine scanning variants	(22)

Table S2 – Comparison of features used in mCSM-PPI1 and mCSM-PPI2. The first version of mCSM-PPI used our graph-based signatures to model the environment of the wild-type residue and pharmacophore to account for the effects of physicochemical changes caused by a point mutation as evidence to train a predictive model. In addition to those two types of features, mCSM-PPI2 also includes six new different classes of features that model different effects of single-point variants.

Type of Feature	Features	Tool	mCSM-PPI1	mCSM-PPI2
Graph-based Signatures	Distance patterns	mCSM (10)	Yes	**Yes**
Pharmacophore Changes	Hydrophobic, Positive, Negative, Hydrogen acceptor, hydrogen donor, aromatic, sulphur and neutral	mCSM (10)	Yes	**Yes**
Wild-type residue Environment	Relative Solvent Accessibility, torsion angle Phi, Residue depth, amino-acid content of chain in which the wild-type resides in percentage: aliphatic, aromatic, positively charged, negatively charged and uncharged	BioPython (11) iFeature (12)	No	**Yes**
Nature of Wild-type and Mutant Residues	Is glycine? Is glycine and has a positive Phi torsion angle? Is proline?	BioPython (11)	No	**Yes**
Evolutionary Information	PSSM Score	Blast 2.2.6	No	**Yes**
Non-Covalent interaction network metrics	Difference between contacts: Van der Waals', aromatic and hydrogen bonds. Complex network metrics for the contact graph of the closest interface of interactions: betweenness, authority score, central points, number of edges. Protein contact potentials scores: SIMK990103, MIYS960101 and ZHAC000104	Arpeggio (13), iGraph, AAindex Database (6)	No	**Yes**
Energetic Terms	Electrostatic interaction between molecules, cost of	FoldX (14)	No	**Yes**

	having a cis peptide bond and Gibbs free energy change			
Atomic Fluctuation	Score using calpha and pfanm force-fields	Bio3D (15)	No	**Yes**

Table S3 - Protein contact potentials scores from the AAindex database (6) used for mCSM-PPI2 as part of its workflow.

AAindex Code	Description	Reference
MIYS960101	Quasichemical energy of transfer of amino acids from water to the protein environment.	(7)
SIMK990103	Distance-dependent statistical potential (contacts within 7.5-10 Å)	(8)
ZHAC000104	Environment-dependent residue contact energies	(9)

Table S4 - mCSM-PPI2 performance comparison on classification using two thresholds (0 and 0.2 kcal/mol) in which ΔΔG >= threshold was considered to increase affinity and ΔΔG < threshold decrease affinity. The performance for each method was calculated over the performance on training. mCSM-PPI2 shows a better and more balanced performance on increasing and decreasing affinity mutations than MutaBind.

		|ΔΔG| > 0 kcal/mol			**|ΔΔG| > 0.2 kcal/mol**		
Method	**Δ Affinity**	**Precision**	**Recall**	**AUC**	**Precision**	**Recall**	**AUC**
mCSM-PPI2	**Increase**	**0.90**	**0.87**	**0.88**	**0.89**	**0.85**	**0.84**
	Decrease	**0.87**	**0.90**	**0.88**	**0.79**	**0.84**	**0.84**
MutaBind	**Increase**	0.62	0.24	0.60*	0.57	0.41	0.66*
	Decrease	0.81	0.96	0.60*	0.84	0.91	0.66*

* $p < 0.05$ by z transformation test compared to mCSM-PPI2.

Table S5 – mCSM-PPI2 training performance comparison across different subsets of variants derived from SKEMPI. The performance of mCSM-PPI2 was calculated after running 10-fold cross-validation 10 times using 90% of the dataset for training and 10% for testing.

SKEMPI	Method	Correlation (ρ)	RMSE (kcal/mol)
S2007	**mCSM-PPI2**	**0.83**	**1.02**
	mCSM-PPI1	0.80*	1.25
	BeAtMuSiC	0.39*	1.81
S1964	**mCSM-PPI2**	**0.82**	**1.08**
	MutaBind	0.78	1.20
	FoldX	0.40*	2.12
	MMPBSA	0.44*	6.45
S1327	**mCSM-PPI2**	**0.80**	**1.12**
	SAAMBE	0.62*	NA
S1102	**mCSM-PPI2**	**0.81**	**1.19**
	iSEE	0.80*	1.41

* $p < 0.05$ by Fisher r-to-z transformation test compared to mCSM-PPI2.

Table S6 – Performance of predictive model across the different types of attributes used to build mCSM_PPI2. For this set of experiments the models were trained and performance evaluated on each type of feature separately. The performance was calculated after running 10-fold cross-validation 10 times using 90% of the dataset S8338 and 10% for testing.

Type of Feature	Correlation (ρ)	RMSE (kcal/mol)
Graph-based signatures + pharmacophores changes	0.57	2.75
Wild-type residue Environment	0.28	3.13
Nature of Wild-type and Mutant Residues	0.13	4.08
Evolutionary Information	0.46	2.84
Non-Covalent interaction network metrics	0.38	3.62
Energetic Terms	0.40	3.53
Atomic Fluctuation	0.11	5.43

Table S7 – Distribution of mutations over protein-protein complexes for the dataset of alanine-scanning experimental mutations.

PDB code	Number of mutations
1A22	64
1GC1	49
1DAN	43
1JRH	31
1BXI	30
1A4Y	28
1VFB	28
3HFM	25
1DFJ	14
1BRS	14
1JCK	9
1F47	9
1AHW	8
1CBW	8
1FCC	8
1DN2	5
1FC2	3
2PTC	1
1NMB	1

FIGURES

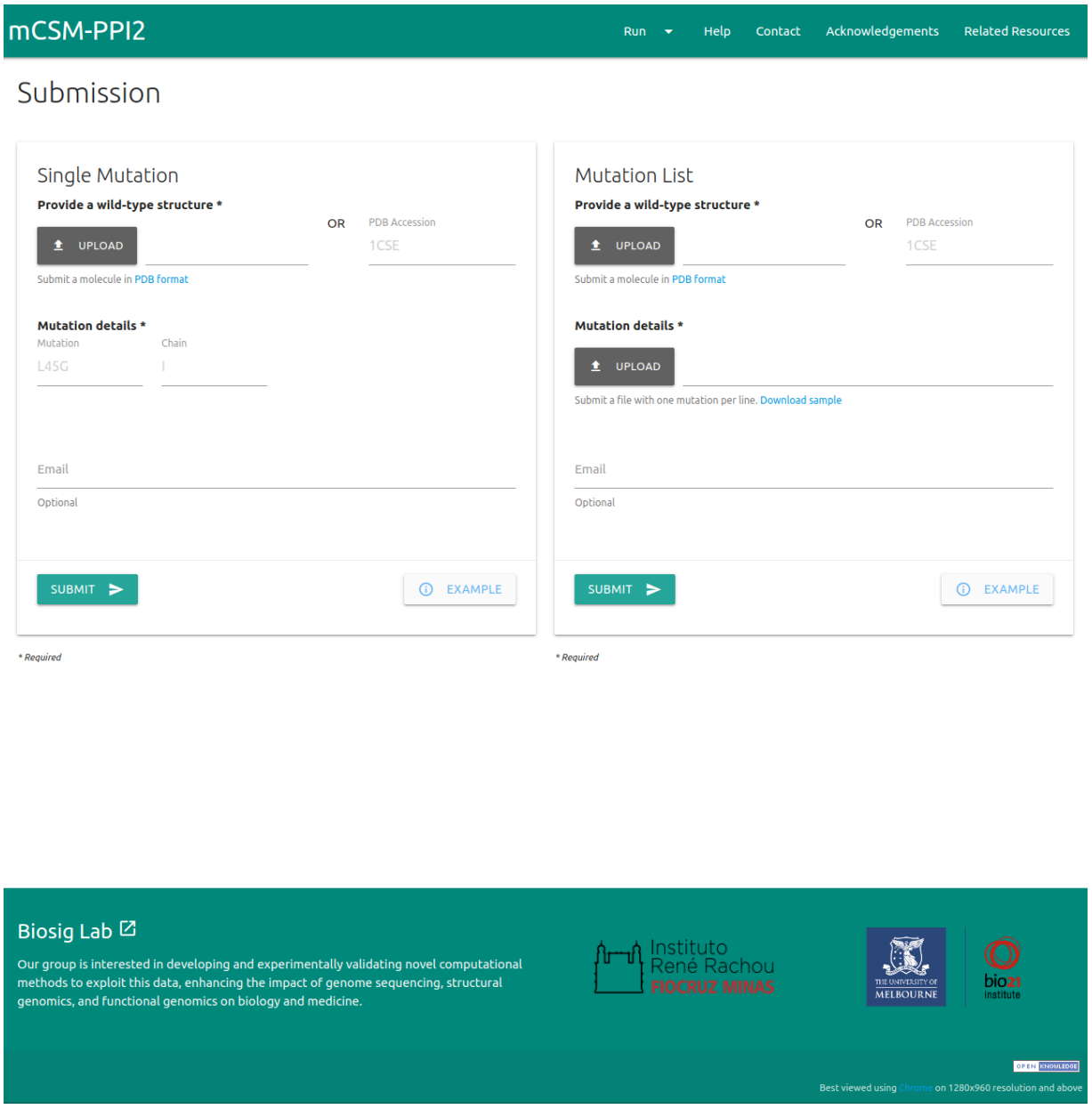


Figure S1 - Submission page for user-specified mutations. Two options are available. For the “Single Mutation” option mCSM-PPI2 requires one to specify a string containing the wild-type, wild-type residue one-letter code, the position in the structure, the mutant residue one-letter code and the chain identifier. For the “Mutation List” option users are asked to provide a file with a list of mutations for batch processing. For both options a PDB file is also required.

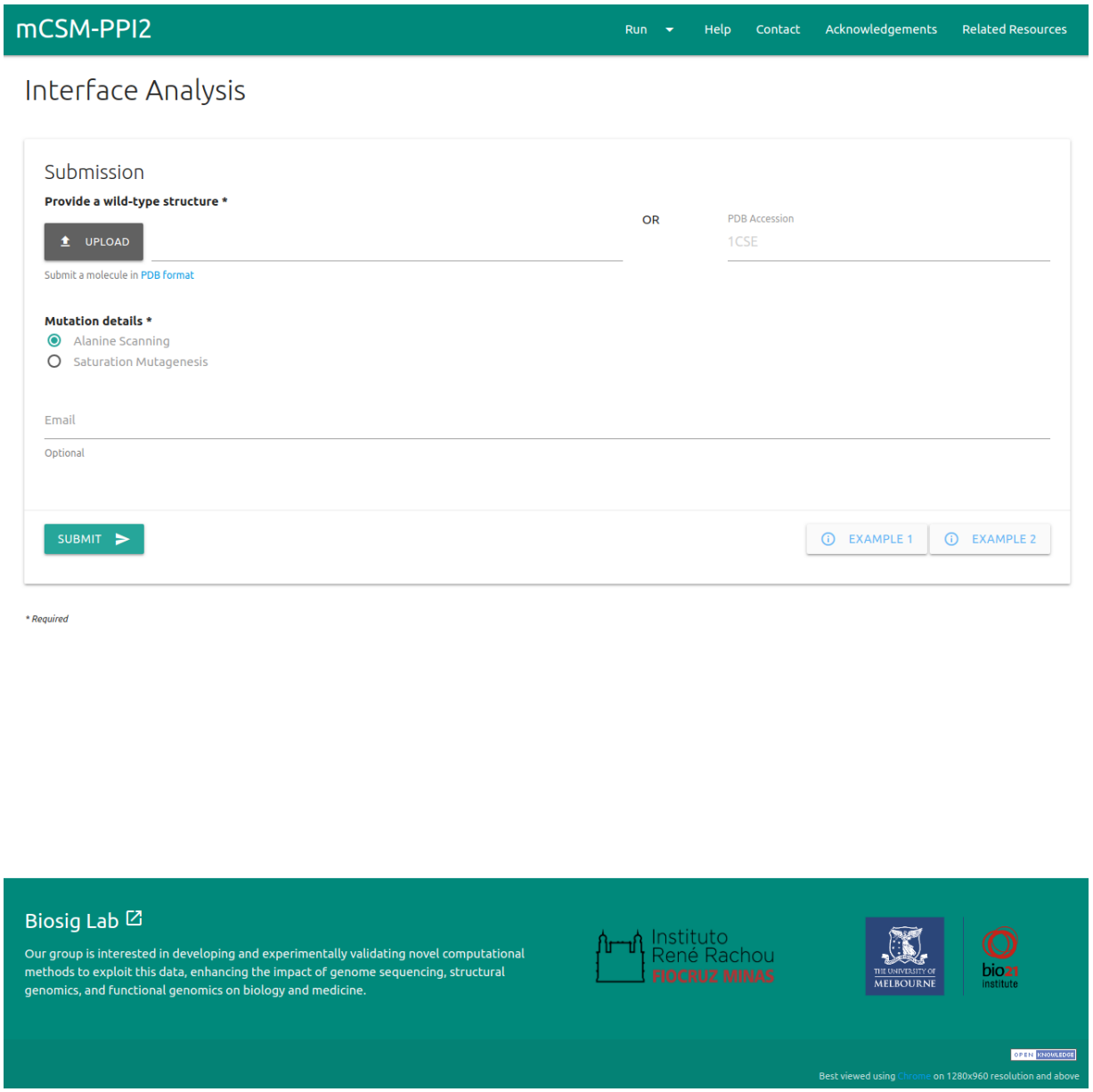


Figure S2 - Input page for Interface analysis. For assessing the effects of mutations at protein-protein interfaces the server requires the user to provide a PDB file or a PDB accession code and select one of two options: alanine scanning (all interface residues are mutated to alanine) or saturation mutagenesis (all interface residues are mutated to every other amino acid).

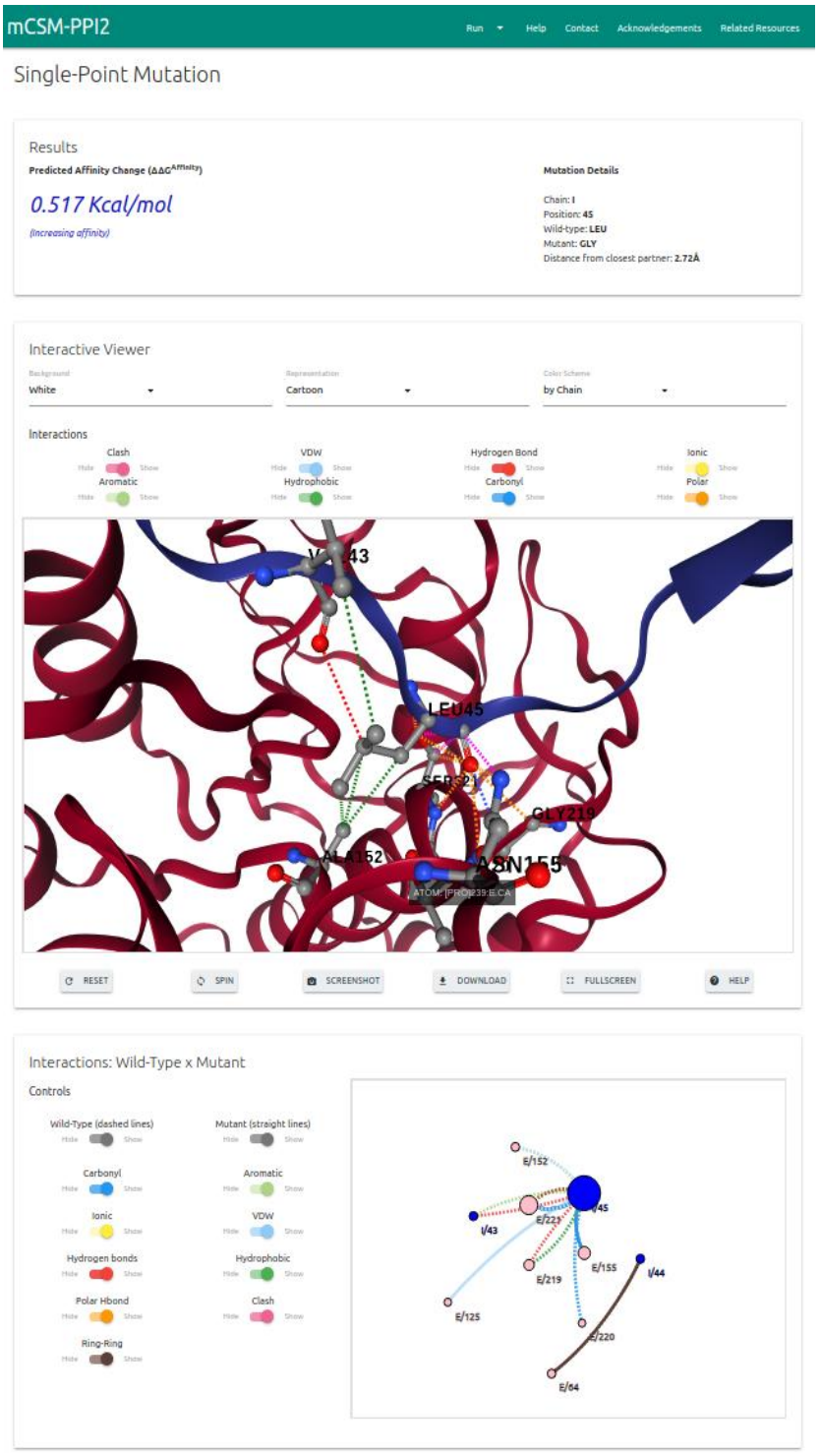


Figure S3 - Results page for “Single Mutation” option. mCSM-PPI2 outputs the change in binding free energy (in kcal/mol) on the top panel alongside with details on the input mutation. An interactive 3D viewer allows for analysis of non-covalent interactions at the position specified on input. Lastly, a 2D graph displays the interactions of wild-type and mutant residues. In both cases controllers are provided in order to hide or show specific interactions.

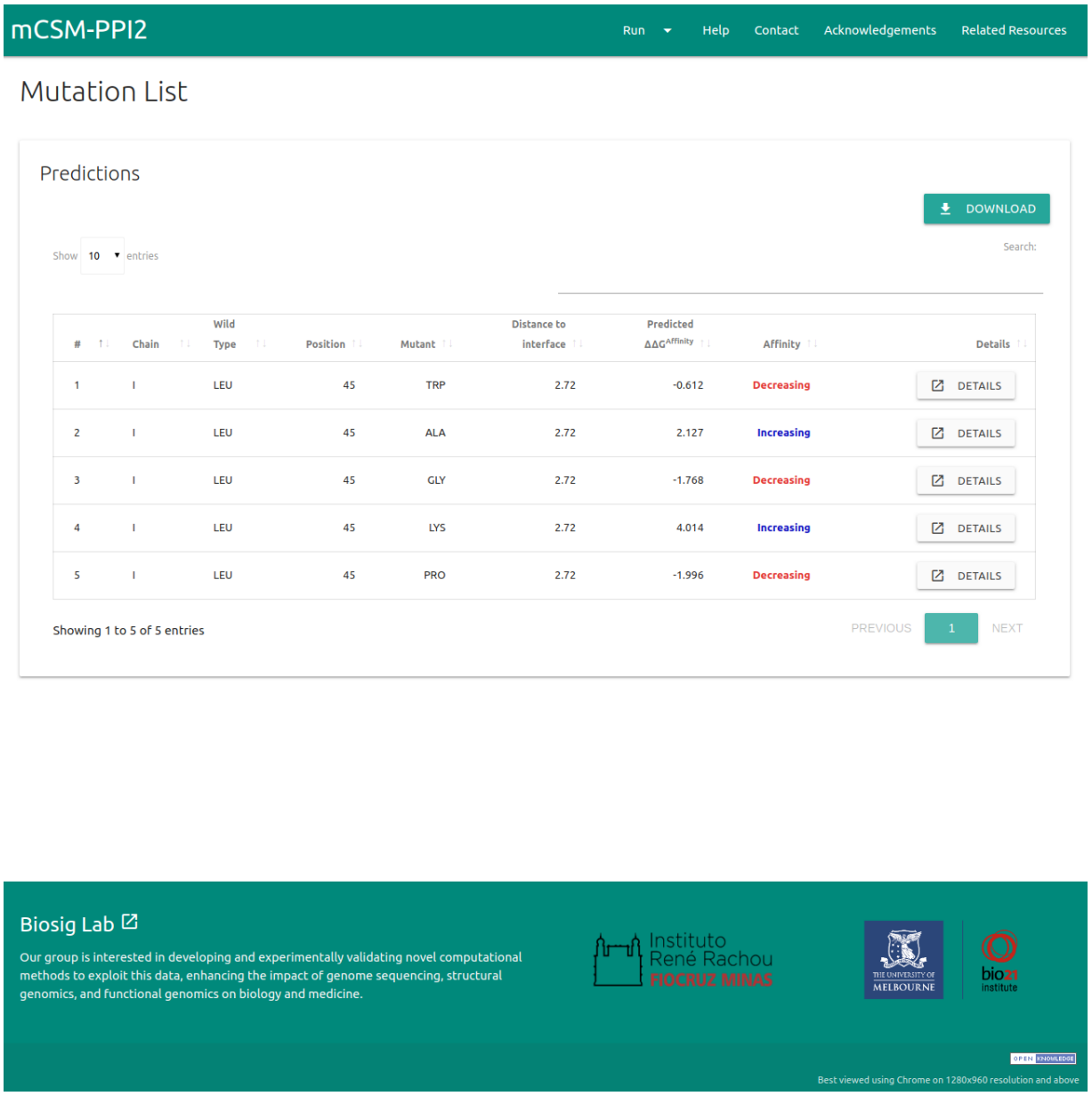


Figure S4 - Results page for “Mutation List” option. The results are summarised in a downloadable table from which users can access details for each single mutation.

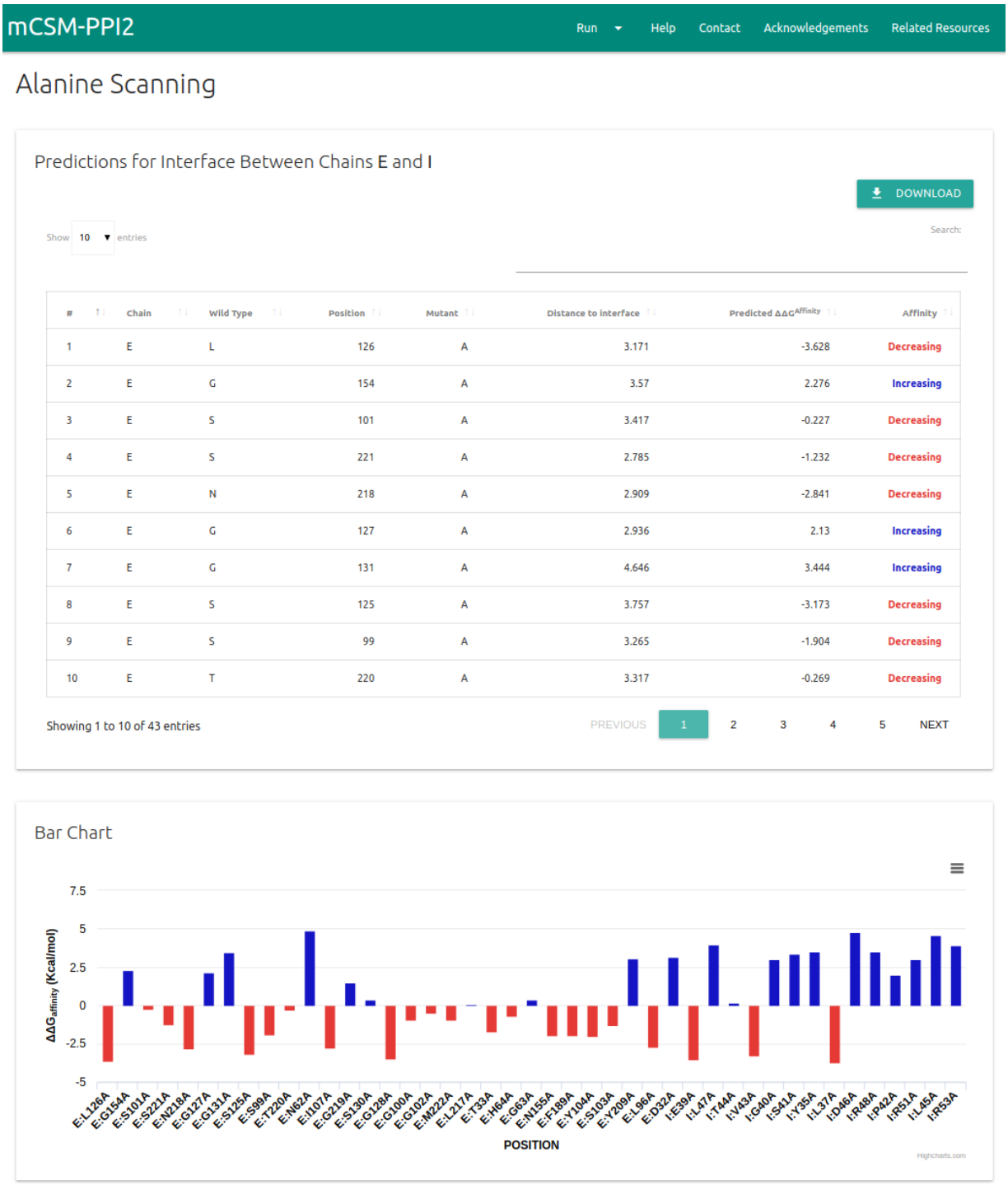


Figure S5 - Table and Bar graph for Alanine Scanning results page. For each interface identified on the protein complex submitted, mCSM-PPI2 shows a downloadable table with all mutations and the results are also summarised in a bar chart in which the bars are coloured according to the predicted change in binding affinity.

Figure S6 - 3D viewer for Alanine Scanning results page. A 3D viewer in which interface residues are coloured according to the predicted change in binding affinity is also shown at the bottom of the results page. A set of controllers are available for customising the structure according to the user's need.

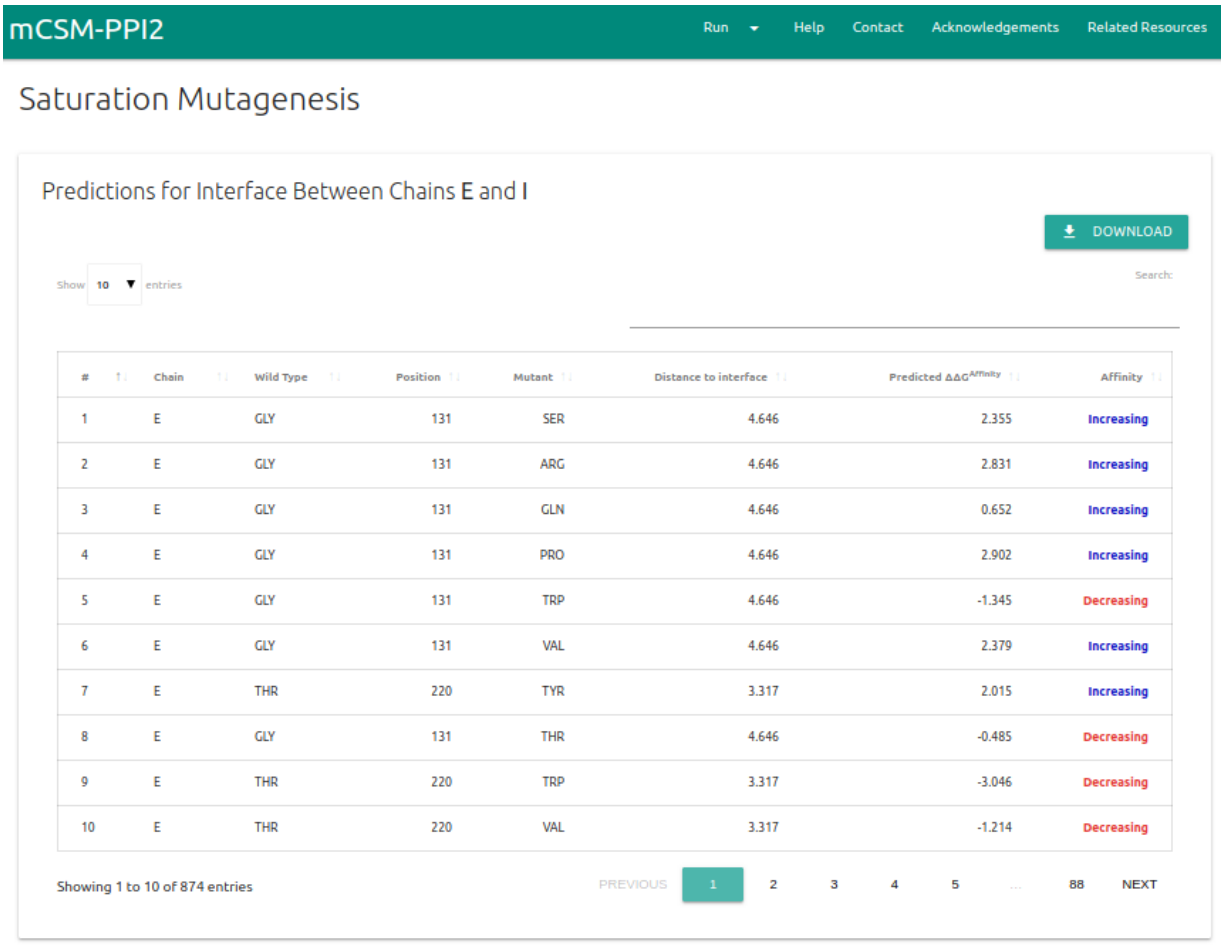


Figure S7 - Table of results for Saturation Mutagenesis option. Similarly to the “Mutation List” and “Alanine Scanning”, for each interface identified on the Saturation mutagenesis option, the results are compiled in a downloadable table.

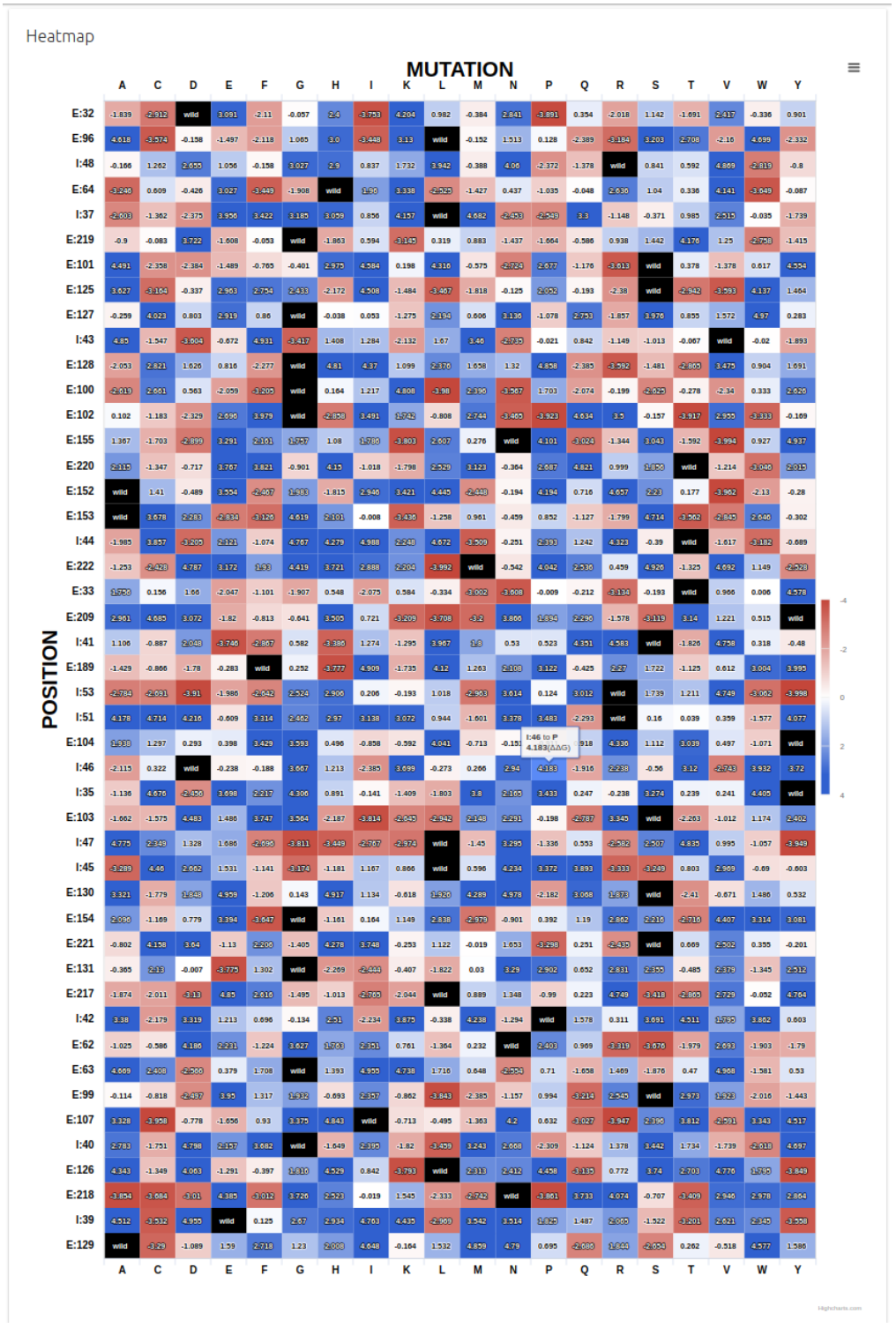


Figure S8 - Heatmap for Saturation Mutagenesis option. The results compiled in the table for saturation mutagenesis are also summarised in a heatmap in which every mutation is coloured according to the predicted effect.

Figure S9 - 3D viewer for Saturation Mutagenesis results page. mCSM-PPI2 also shows an interactive 3D viewer for the saturation mutagenesis option in which the interface residues are coloured according to the average predicted change in binding affinity.

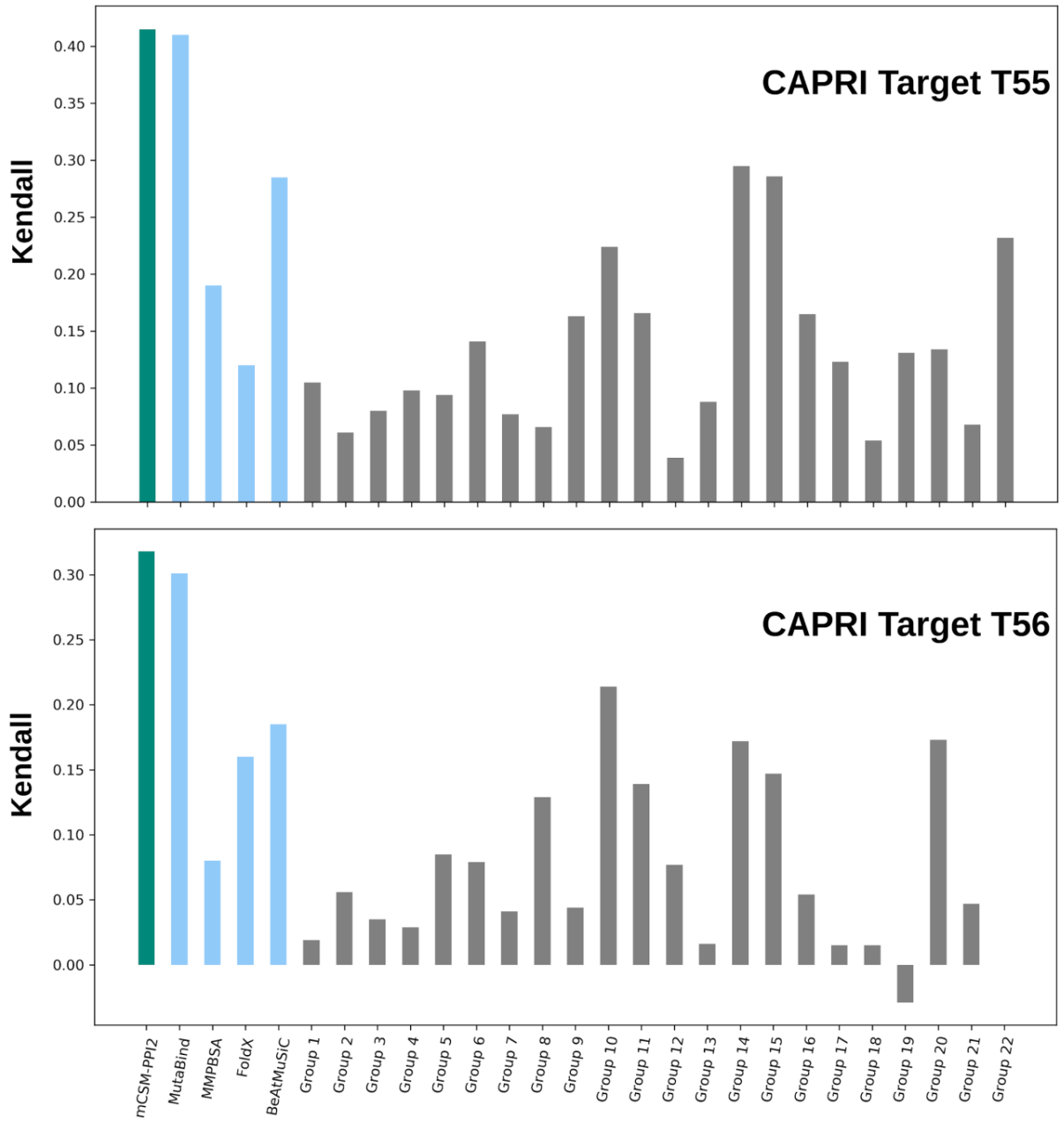


Figure S10 - Performance comparison on CAPRI round 26th. mCSM-PPI2 (green) outperforms all other 26 methods with a Kendall's coefficient of 0.42 and 0.32 for datasets based on Targets T55 and T56, respectively. Bar coloured as blue indicate methods that used this same dataset as a benchmark on their studies. Other methods which participated on CAPRI by the time the dataset was released in 2012 were coloured in grey.

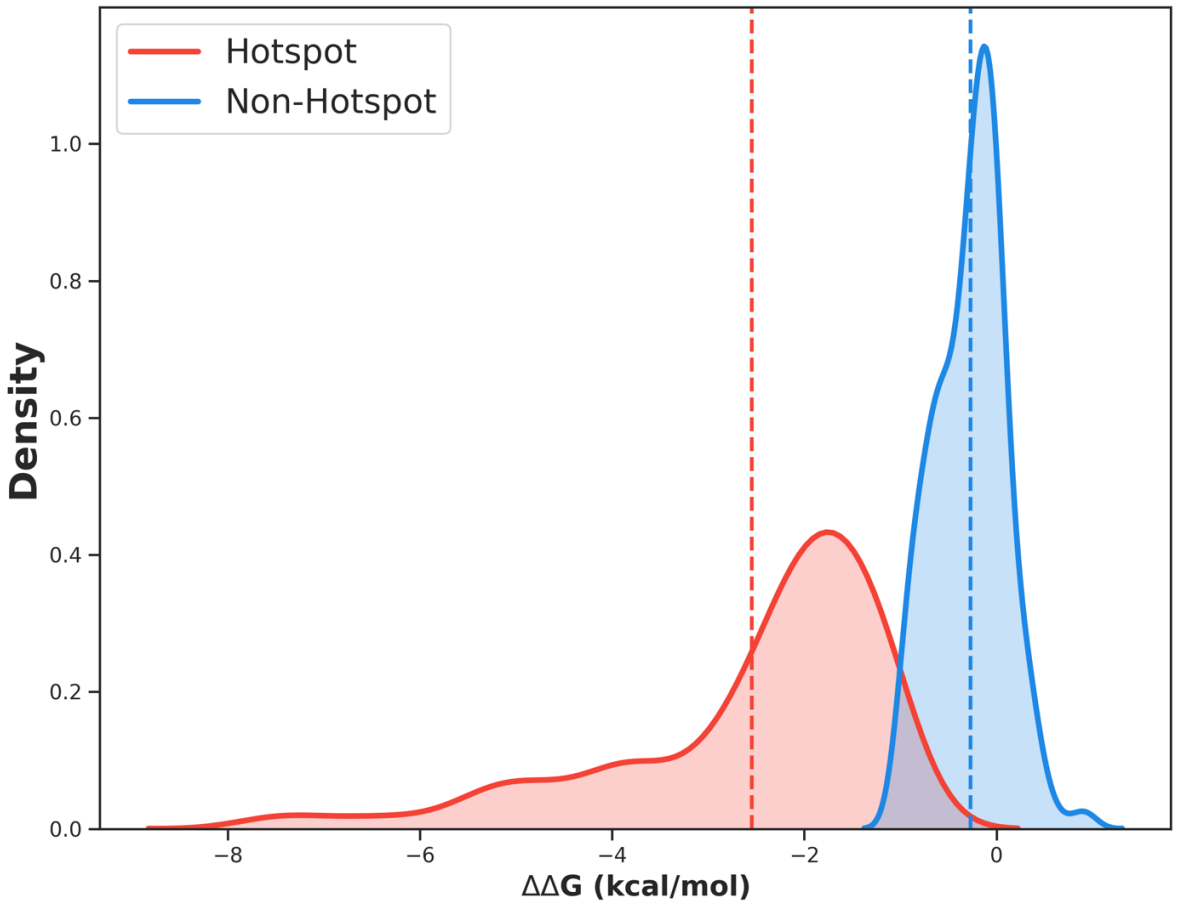


Figure S11 - Density distribution of ΔΔG predictions by mCSM-PPI2 for mutations that were experimentally assigned as hotspots. While the majority of neutral mutations (blue) were predicted to have little impact, most of hotspots mutations (red) were predicted to have a significant decrease in the binding affinity of the complex.

REFERENCES

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2012) *Scikit-learn: Machine Learning in Python*.
2. Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely randomized trees. *Machine Learning*, **63**, 3-42.
http://dx.doi.org/10.1007/s10994-006-6226-1

3. Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779-815.
http://www.ncbi.nlm.nih.gov/pubmed/8254673
http://dx.doi.org/10.1006/jmbi.1993.1626

4. Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.M., Wilson, I.A. and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816-821.
http://www.ncbi.nlm.nih.gov/pubmed/21566186
http://dx.doi.org/10.1126/science.1202617

5. Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol*, **30**, 543-548.
http://www.ncbi.nlm.nih.gov/pubmed/22634563
http://dx.doi.org/10.1038/nbt.2214

6. Kawashima, S. and Kanehisa, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*, **28**, 374.
http://www.ncbi.nlm.nih.gov/pubmed/10592278

7. Miyazawa, S. and Jernigan, R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, **256**, 623-644.
http://www.ncbi.nlm.nih.gov/pubmed/8604144
http://dx.doi.org/10.1006/jmbi.1996.0114

8. Koehl, P. and Levitt, M. (2002) Improved recognition of native-like protein structures using a family of designed sequences. *Proc Natl Acad Sci U S A*, **99**, 691-696.
http://www.ncbi.nlm.nih.gov/pubmed/11782533
http://dx.doi.org/10.1073/pnas.022408799

9. Zhang, C. and Kim, S.H. (2000) Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci U S A*, **97**, 2550-2555.
http://www.ncbi.nlm.nih.gov/pubmed/10706611
http://dx.doi.org/10.1073/pnas.040573597

10. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335-342.
http://www.ncbi.nlm.nih.gov/pubmed/24281696
http://dx.doi.org/10.1093/bioinformatics/btt691

11. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422-1423.
http://www.ncbi.nlm.nih.gov/pubmed/19304878
http://dx.doi.org/10.1093/bioinformatics/btp163

12. Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.C. *et al.* (2018) iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34**, 2499-2502.
http://www.ncbi.nlm.nih.gov/pubmed/29528364
http://dx.doi.org/10.1093/bioinformatics/bty140

13. Jubb, H.C., Higueruelo, A.P., Ochoa-Montano, B., Pitt, W.R., Ascher, D.B. and Blundell, T.L. (2017) Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J Mol Biol*, **429**, 365-371.
http://www.ncbi.nlm.nih.gov/pubmed/27964945
http://dx.doi.org/10.1016/j.jmb.2016.12.004

14. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res*, **33**, W382-388.
http://www.ncbi.nlm.nih.gov/pubmed/15980494
http://dx.doi.org/10.1093/nar/gki387

15. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695-2696.
http://www.ncbi.nlm.nih.gov/pubmed/16940322
http://dx.doi.org/10.1093/bioinformatics/btl461

16. Jankauskaite, J., Jimenez-Garcia, B., Dapkunas, J., Fernandez-Recio, J. and Moal, I.H. (2019) SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, **35**, 462-469.
http://www.ncbi.nlm.nih.gov/pubmed/30020414
http://dx.doi.org/10.1093/bioinformatics/bty635

17. Moal, I.H. and Fernandez-Recio, J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, **28**, 2600-2607.
http://www.ncbi.nlm.nih.gov/pubmed/22859501
http://dx.doi.org/10.1093/bioinformatics/bts489

18. Dehouck, Y., Kwasigroch, J.M., Rooman, M. and Gilis, D. (2013) BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res*, **41**, W333-339.
http://www.ncbi.nlm.nih.gov/pubmed/23723246
http://dx.doi.org/10.1093/nar/gkt450

19. Li, M., Simonetti, F.L., Goncearenco, A. and Panchenko, A.R. (2016) MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. *Nucleic Acids Res*, **44**, W494-501.
http://www.ncbi.nlm.nih.gov/pubmed/27150810
http://dx.doi.org/10.1093/nar/gkw374

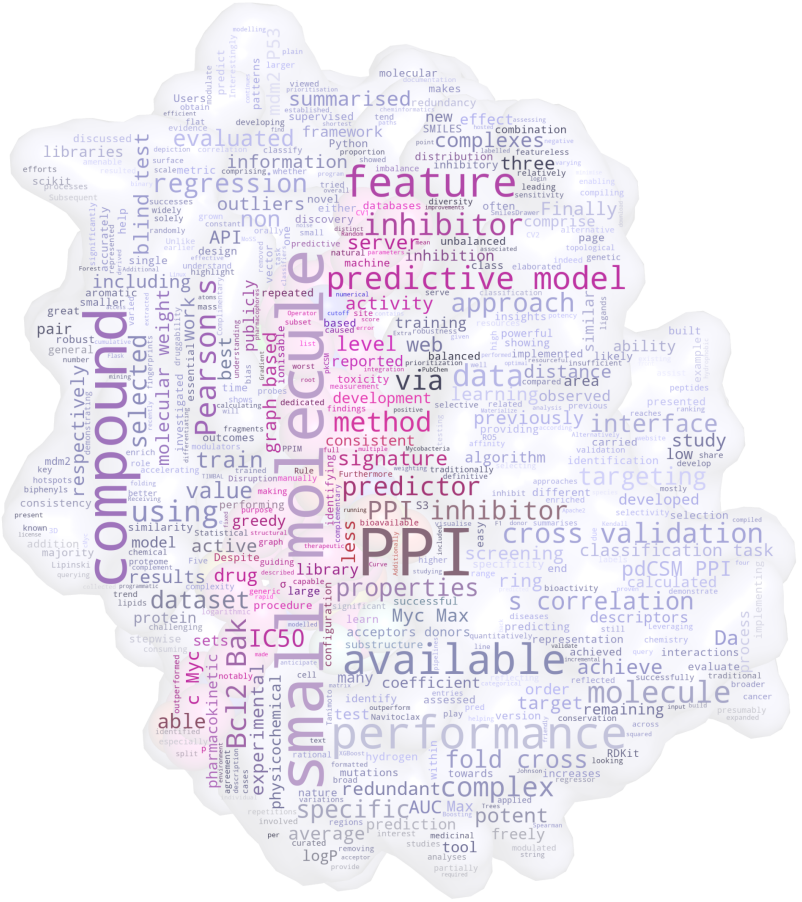
20. Petukh, M., Dai, L. and Alexov, E. (2016) SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations. *Int J Mol Sci*, **17**, 547.
http://www.ncbi.nlm.nih.gov/pubmed/27077847
http://dx.doi.org/10.3390/ijms17040547

21. Geng, C., Vangone, A., Folkers, G.E., Xue, L.C. and Bonvin, A. (2019) iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins*, **87**, 110-119.
http://www.ncbi.nlm.nih.gov/pubmed/30417935
http://dx.doi.org/10.1002/prot.25630

22. Kortemme, T. and Baker, D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, **99**, 14116-14121.
http://www.ncbi.nlm.nih.gov/pubmed/12381794
http://dx.doi.org/10.1073/pnas.202485799

Chapter 5

Using Graph-based Signatures to Identify PPI Inhibitors



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

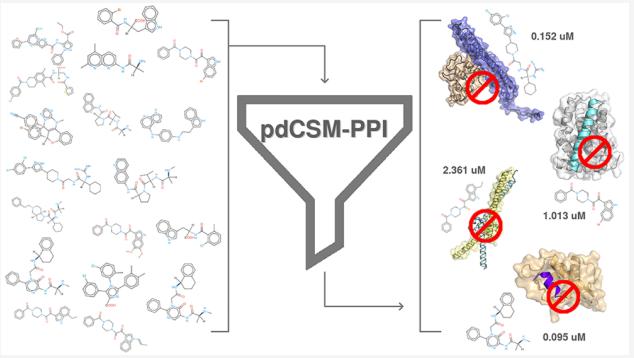
pdCSM-PPI: Using Graph-Based Signatures to Identify Protein–Protein Interaction Inhibitors

Carlos H. M. Rodrigues, Douglas E. V. Pires,* and David B. Ascher*

Cite This: https://doi.org/10.1021/acs.jcim.1c01135 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Protein–protein interactions are promising sites for development of selective drugs; however, they have generally been viewed as challenging targets. Molecules targeting protein–protein interactions tend to be larger and more lipophilic than other drug-like molecules, mimicking the properties of interacting interfaces. Here, we propose a machine learning approach that uses a graph-based representation of small molecules to guide identification of inhibitors modulating protein–protein interactions, pdCSM-PPI. This approach was applied to 21 different PPI targets. We developed interaction-specific models that were able to accurately identify active compounds achieving MCC and F1 scores up to 1, and Pearson's correlations up to 0.87, outperforming previous approaches. Using insights from these individual models, we developed a generic protein–protein interaction modulator predictive model, which accurately predicted IC50 with a Pearson's correlation of 0.64 on a low redundancy blind test. Importantly, we were able to accurately identify active from inactive compounds, achieving an AUC of 0.77 and sensitivity and specificity of 76% and 78%, respectively. We believe pdCSM-PPI will be an important tool to help guide more efficient screening of new PPI inhibitors; it is freely available as an easy-to-use web server and API at http://biosig.unimelb.edu.au/pdcsm_ppi.



INTRODUCTION

Protein–protein interactions (PPIs) play an essential role in most key processes within the cell.[1−4] Disruption of PPIs, often caused by genetic mutations,[5,6] are involved in the development of many diseases, including cancer.[7] Unlike traditional active site drug targets that often share significant conservation with other areas of the proteome leading to off-target effects, the broad diversity of PPIs is of great interest for the development and discovery of selective and specific drugs. PPI interfaces were traditionally viewed as large, flat, and relatively featureless,[8] making them not amenable to small molecule inhibition. Subsequent work, however, has shown that many interfaces can indeed be successfully modulated by small molecules, especially through targeting hotspots at the interface.[9] This has been reflected in the successful design of inhibitors of PPIs, most notably the development of small-molecule modulators of the Bcl2/Bak complex.[10]

Despite these successes, developing PPI inhibitors is still a challenging and time-consuming process.[11,12] Several studies focusing on chemical properties of known PPI inhibitors to streamline the process of screening for new compounds have been developed. These involved a range of machine learning (ML) algorithms applied to limited data sets, including decision trees to model chemical descriptors derived from 25 PPI inhibitors[13] and support vector machines (SVM) applied to 40 distinct compounds.[14] Since then, many efforts to increase data availability for small molecules targeting PPIs have been proposed, most notably 2P2I-DB[15] with over 274 inhibitors targeting 27 different PPIs and iPPI-DB[16] with over 2378 inhibitors acting on 46 PPI families. The emergence of large collections of data prompted the development of more robust data-driven methods such as PPIMpred[17] and SMMPPI.[18] The former used 11 chemical descriptors with SVM to build predictive models to identify inhibitors for three oncogenic PPIs: Bcl2/Bak, mdm2/P53, and c-Myc/Max. More recently, SMMPPI was developed as a two-stage classification workflow based on chemical structure fingerprints, which at first identifies small molecules more likely to inhibit PPIs and on stage 2 attempts to classify whether the compounds have inhibitory activity against 11 different PPI complexes.

The ability to accurately predict compounds more likely to target PPI interfaces can help understand better the complexity and druggability of these regions. We have previously shown that graph-based signature representation of small molecules is a

Received: September 15, 2021

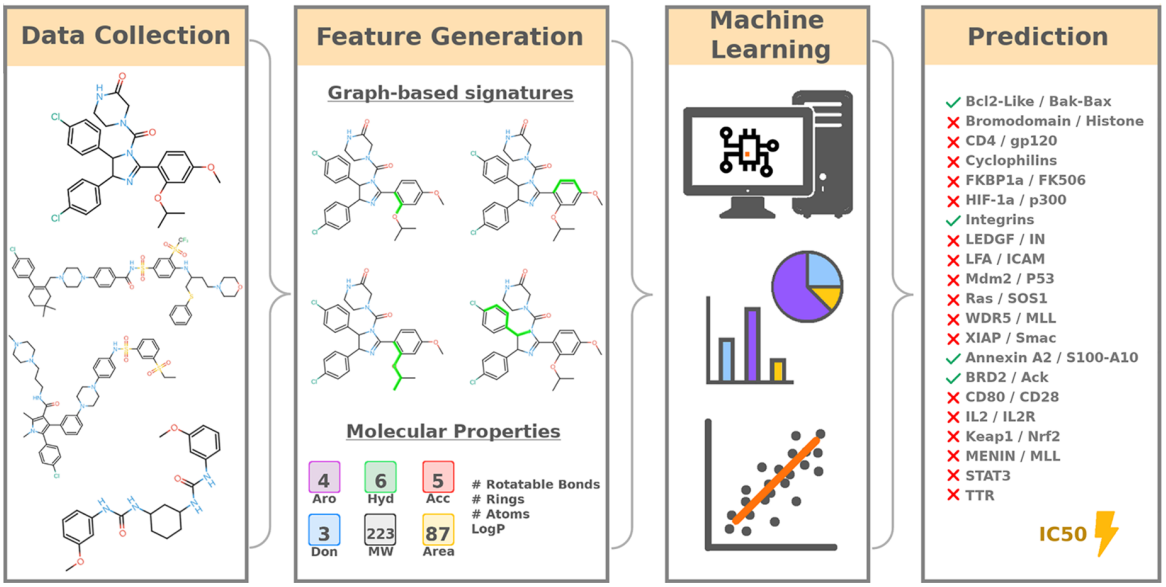


Figure 1. Methodology workflow of pdCSM-PPI. The development of pdCSM-PPI can be divided into four main steps. First, data on small molecules with activity against PPIs, including IC50 values, were collected from the literature. In the second step, geometrical and physicochemical properties for each compound were generated in the form of graph-based signatures. Additional molecular properties such as number of rotatable bonds, rings, number of atoms, LogP value, and fragment-based descriptors, calculated via RDKit, were also included. These were then used to train and validate supervised learning predictive models. Best performing models were selected and made available via a user-friendly web server and API for easy integration with other analysis pipelines.

Table 1. Performances of 21 Target-Specific Predictive Models[a]

	10-fold CV							Blind test						
PPI target	MCC	F1	AUC	TP	TN	FP	FN	MCC	F1	AUC	TP	TN	FP	FN
Bcl2-Like/Bak-Bax	0.94	0.97	0.98	184	192	2	10	0.99	0.99	1.00	65	64	1	0
Bromodomain/Histone	0.83	0.92	0.97	400	383	44	27	0.74	0.86	0.88	138	110	34	5
CD4/gp120	1.00	1.00	1.00	60	60	0	0	0.91	0.95	0.95	18	20	0	2
Cyclophilins	1.00	1.00	1.00	54	54	0	0	0.90	0.95	1.00	19	17	2	0
FKBP1a/FK506	1.00	1.00	1.00	67	67	0	0	0.92	0.96	1.00	23	21	2	0
HIF-1a/p300	0.97	0.98	0.99	60	60	1	1	0.91	0.95	0.98	20	20	1	1
Integrins	0.95	0.98	0.99	700	698	19	17	0.93	0.97	0.99	237	226	14	3
LEDGF/IN	0.87	0.93	0.97	91	78	13	0	0.77	0.87	0.90	23	31	0	8
LFA/ICAM	1.00	1.00	1.00	112	112	0	0	0.95	0.97	0.99	37	38	1	1
Mdm2-Like/P53	0.96	0.98	0.99	326	331	5	10	0.94	0.97	1.00	110	110	4	3
Ras/SOS1	0.95	0.98	0.99	41	41	1	1	0.87	0.93	0.99	14	12	2	0
WDR5/MLL	0.94	0.97	0.97	32	31	1	1	0.92	0.96	0.94	10	12	0	1
XIAP/Smac	0.98	0.99	0.99	132	134	0	3	0.89	0.94	0.94	40	45	0	5
Annexin A2/S100-A10	0.97	0.98	0.98	28	29	0	1	–	–	–	–	–	–	–
BRD2/Ack	1.00	1.00	1.00	32	33	0	0	–	–	–	–	–	–	–
CD80/CD28	1.00	1.00	1.00	35	35	0	0	–	–	–	–	–	–	–
IL2/IL2R	1.00	1.00	1.00	19	19	0	0	–	–	–	–	–	–	–
Keap1/Nrf2	0.96	0.98	0.98	26	28	0	1	–	–	–	–	–	–	–
MENIN/MLL	1.00	1.00	1.00	32	33	0	0	–	–	–	–	–	–	–
STAT3	1.00	1.00	1.00	21	21	0	0	–	–	–	–	–	–	–
TTR	1.00	1.00	1.00	35	35	0	0	–	–	–	–	–	–	–

[a]Results are shown in terms of Matthew's correlation coefficient (MCC), F1 score, area under the ROC curve (AUC), true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

powerful tool for providing novel insights into, and accurately predicting, small molecule pharmacokinetic, toxicity, and bioactivity properties.[19−22] Here, we have explored the application of this approach to the discovery of PPI inhibitors (Figure 1). We have made our method, pdCSM-PPI, freely available to assist with ongoing PPI inhibitor screening efforts.

RESULTS AND DISCUSSION

Common Properties of PPI Inhibitors. Most compounds in our data set show physicochemical properties in agreement with the widely used Lipinski Rule of Five (RO5) for orally bioavailable drugs, with an average molecular weight of 483 Da, four hydrogen acceptors, three donors, and logP of 3.90 (Figure S1), presumably reflecting a natural bias in the screening

https://doi.org/10.1021/acs.jcim.1c01135
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

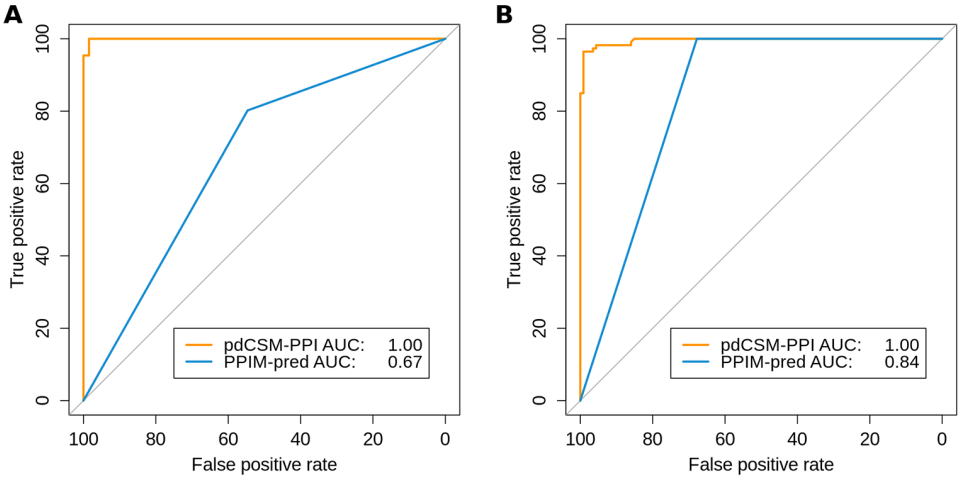


Figure 2. Performance comparison of pdCSM-PPI with PPIM-pred. Both models were evaluated on their ability to distinguish between inhibitors and noninhibitors of two distinct PPIs: (A) Bcl2/Bak and (B) mdm2/P53. pdCSM-PPI (orange) outperforms PPIM-pred (blue) in both cases (p-value <0.05).

libraries used to obtain these molecules. Interestingly, more potent inhibitors (IC50 < 1 μm) presented a small deviation to RO5 in terms of molecular weight, indicating that these could have been elaborated/grown through medicinal chemistry to achieve higher selectivity toward specific PPI interfaces (average molecular weight of 513 Da, logP of 3.97, five hydrogen acceptors, and three hydrogen donors). In fact, a certain degree of flexibility to the rule is observed for small molecules within our data set; for example, Navitoclax, a known potent inhibitor for the Bcl2/Bak complex, has a molecular mass of 973 Da, 14 acceptors, two donors, and 9.6 logP. Furthermore, the most potent molecules are enriched with complex ring substructures (Figure S2), including biphenyls which have been previously shown to be related to high levels of specificity.[23] These findings are also consistent with previous studies showing that PPI inhibitors tend to have more aromatic rings than other ligands.[24,25]

Identifying PPI-Specific Inhibitors. The identification of chemical compounds that modulate PPIs to serve as probes for guiding rational design of new and more potent small molecule inhibitors is essential for accelerating drug discovery. Many efforts have been implemented into compiling this information into large publicly available databases, from libraries with a broader range of small molecules, lipids and peptides,[26,27] to resources dedicated solely to compounds targeting PPIs.[15,16,28]

In this study, we manually curated data for small molecules with experimental inhibitory activity for 21 distinct PPI targets in Simplified Molecular-Input Line-Entry System (SMILES) format. In order to complement and expand on previous works, here we generated two sets of features for each compound, namely, graph-based signatures[19] and physicochemical descriptors calculated via RDKit descriptors, which were then used as evidence to train 21 different supervised learning predictive models, one for each PPI target. Feature selection was carried out via a greedy stepwise approach, and the best performing models were selected. Performances on the training set under 10-fold cross-validation for all methods were given in terms of F1 score, area under the ROC curve (AUC) and Matthews correlation coefficient (MCC) and are summarized in Table 1 and Figures S3−S5.

As each model was trained individually, the optimal number of features selected across all predictive models varied from 3 to 65. However, we observed that the best performing models used three main types of features which capture critical aspects of PPI inhibitors extensively described in the literature: (1) Molecular surface area descriptors (MOE), as evidenced by previous studies, surface properties are key elements of more efficient PPI inhibitors;[29] (2) Distance patterns from graph-based signatures involving atoms on aromatic ring structures emphasize the role aromatic rings play on these types of compounds[28] and are consistent with the findings of our substructure mining presented previously; (3) Fragment descriptors capture common structural patterns for each PPI target (the number of bicyclic fragments for example). Furthermore, analysis of feature importance values output from the ExtraTrees algorithm shows a balanced distribution of importance across all classifiers, indicating a synergistic contribution by all features for the final prediction (Tables S1−21, Supporting Information).

The performances of each target-specific predictive model were further evaluated on nonredundant blind test sets, and outcomes were compared with those reported for PPIM-pred[17] and SMMPPI.[18] PPIM-pred is a ML-based predictor for inhibitors targeting three distinct PPIs: Bcl2/Bak, Mdm2/P53, and c-Myc/Max. However, as machine learning is a data-driven approach highly dependent on the amount of data used for training a predictive model, and given the lack of data available in our data set after removing redundancy for inhibitors of c-Myc/Max (eight compounds), we are only able to compare the results of two of our predictive models with PPIM-pred, namely, Bcl2/Bak and Mdm2/P53. For a fair comparison between the two methods, here we removed from our training set all compounds with a Tanimoto similarity score greater than 0.8 from the test set reported by PPIM-pred. As our method has been trained on a larger data set of inhibitors for these two PPI targets using a more diverse set of features, it is no surprise that our approach significantly outperformed the metrics reported by PPIM-pred (p-value < 0.05), showing a more balanced prediction for inhibitors on both targets (Figure 2 and Table S22).

SMMPPI is a more recent ML-based method built on a more robust data set of PPI inhibitors. The method proposes class-specific predictors of inhibitors for 11 clinically important PPI families, nine of which are included among those developed using our approach. Here, we used the same test sets available on the SMMPPI study for seven PPI targets, and test sets for the

https://doi.org/10.1021/acs.jcim.1c01135
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Performance of Regression Models Predicting IC50 Values on Training under 10-Fold Cross-Validation and Nonredundant Blind Test for Inhibitors of Nine Different PPI Targets[a]

	10-fold CV					Blind test				
PPI target	ρ	τ	r_s	RMSE	MAE	ρ	τ	r_s	RMSE	MAE
BCL2-Like/Bax-Bak	0.76	0.59	0.79	0.73	0.54	0.64	0.43	0.63	0.79	0.65
Bromodomain/Histone	0.68	0.46	0.63	0.88	0.69	0.44	0.34	0.51	0.84	0.61
Cyclophilins	0.91	0.60	0.81	0.55	0.41	0.87	0.48	0.67	0.57	0.39
HIF-1a/p300	0.51	0.36	0.57	0.90	0.63	0.28	0.09	0.13	0.83	0.62
Integrins	0.59	0.39	0.55	1.26	0.99	0.49	0.33	0.47	1.31	1.00
LEDGF/IN	0.33	0.21	0.31	0.52	0.41	−0.03	−0.02	−0.09	0.60	0.52
LFA/ICAM	0.71	0.48	0.65	0.69	0.50	0.67	0.37	0.59	0.85	0.65
Mdm2-Like/P53	0.76	0.56	0.74	0.79	0.60	0.63	0.45	0.61	0.76	0.61
XIAP/Smac	0.67	0.29	0.40	0.75	0.52	0.44	0.37	0.50	0.89	0.57

[a]Results are shown in terms of coefficients of correlation, namely, Pearson (ρ), Kendall (τ), Spearman (r_s), and RMSE.

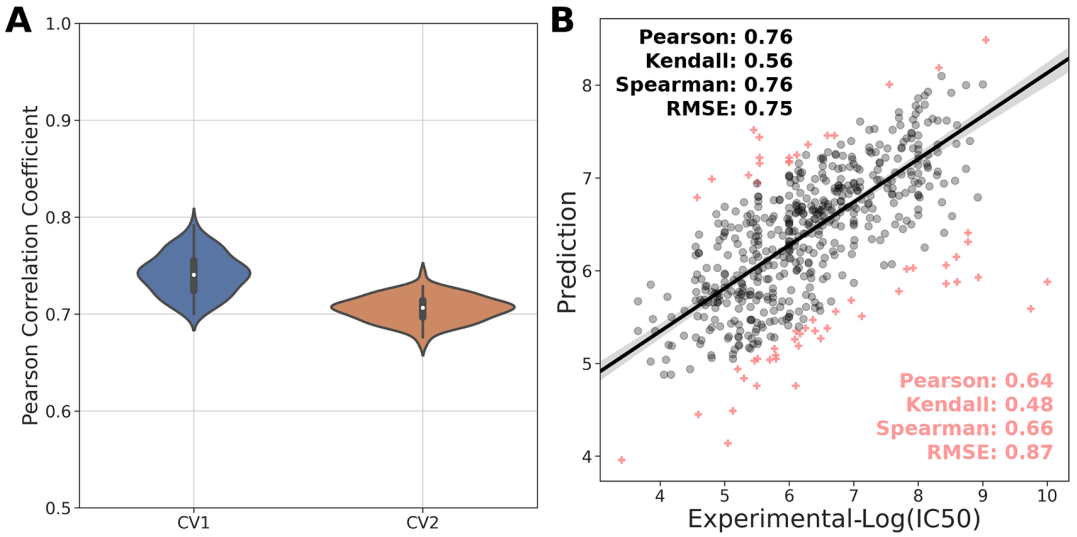


Figure 3. Performance evaluation of general predictor of PPI inhibitor potency. (A) Summary of performances on CV1 (80/20 split) and CV2 (50/50 split) for a general predictor in terms of Pearson's Correlation. (B) Distribution of experimental inhibitory activity versus predictions on a blind test (low redundant) at molecule similarity level of 0.80. IC50 values are shown on a log scale.

WDR5/MLL1 and CD4/gp120 were used from our 75/25 split as these were not available on SMMPPI. For the sake of comparison, all small molecules with similarity greater than 0.8 were removed from our training set. Despite using a smaller number of features for all predictors, our approach achieves similar performance to those reported for SMMPPI for inhibitors on all nine PPI targets in terms of MCC, F1 score, and AUC (Table S23).

Finally, the 21 class-specific predictive models were used to scan the NCI database with over 224,000 compounds, and results are available at http://biosig.unimelb.edu.au/pdcsm_ppi/data

Predicting PPI Inhibitor Potency. After demonstrating the ability of our approach to classify molecules as either inhibitors or noninhibitors, we investigated whether we could predict the ability of small molecules to inhibit PPIs quantitatively, as a regression task. New predictive models were trained and evaluated using half maximal inhibitory concentration (IC50), in a logarithmic scale, as outcomes. Here, we were able to retrieve IC50 values for 3972 inhibitors from TIMBAL and iPPI-DB targeting 45 different PPIs. In order to minimize redundancy, for each PPI target, we clustered compounds with 0.8 Tanimoto similarity and selected one small molecule per cluster. Similar to our criteria used for the classifiers, here we chose nine PPI targets where the number of small molecules after clustering was greater than 40 to proceed with the regression task: Bcl2-Like/Bax-Bak, Bromodomain/Histone, Cyclophilins, HIF-1a/p300, Integrins, LEDGF/IN, LFA/ICAM, Mdm2-Like/P53, and XIAP/Smac. A summary of the distribution of compounds across the 45 different PPI targets, before and after clustering, is shown in Table S24. Data sets were then split into 75% for training and 25% as a nonredundant blind test using the same strategy previously described based on the number of compounds per cluster. Here, the predictive models were trained using the same set of features selected during our classification task for each PPI target.

Performances on training under 10-fold cross-validation ranged from Pearson's and Kendall's correlations of 0.33 and 0.21 (RMSE = 0.52) for inhibitors of the LEDGF/IN target to Pearson's and Kendall's correlations of 0.91 and 0.60 (RMSE = 0.55) for Cyclophilins (Table 2). Surprisingly, the predictive model for inhibitors targeting Integrins achieved Pearson's and Kendall's correlations under 10-fold cross-validation of 0.59 and 0.39 (RMSE = 0.55), respectively, despite having the largest number of entries available for training (680). On the nonredundant blind test, the results remained consistent with Pearson's and Kendall's correlations of 0.49 and 0.33 (RMSE = 1.31), respectively. Even though, the classifier for this PPI target

https://doi.org/10.1021/acs.jcim.1c01135
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

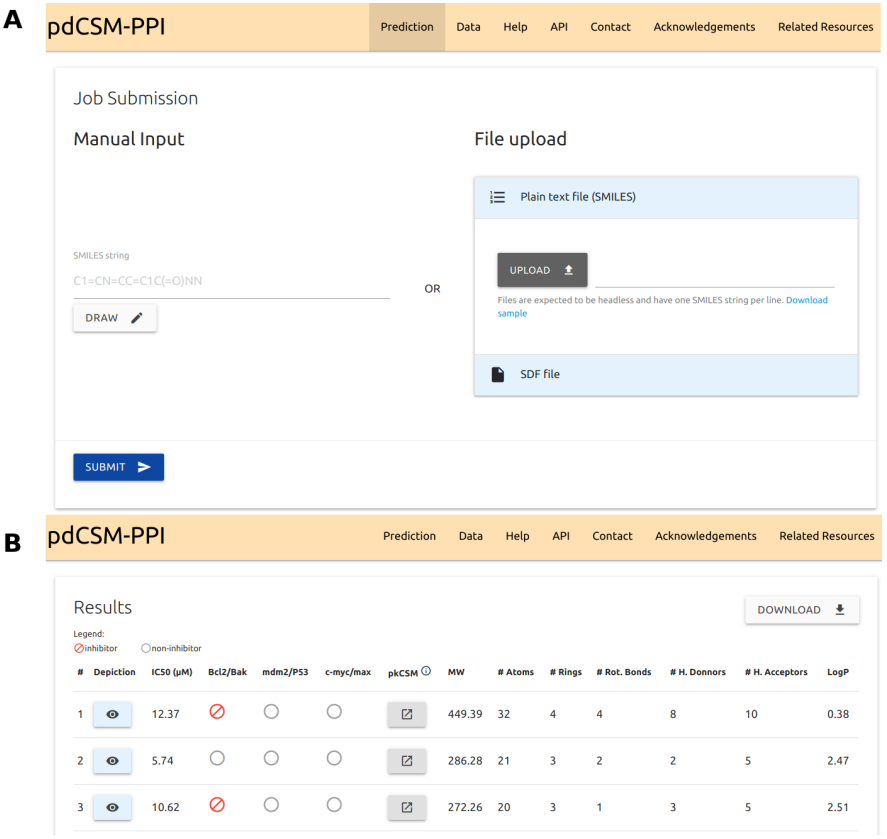


Figure 4. pdCSM-PPI web interface. (A) The submission page allows users to submit one single molecule as a SMILES string or upload a list of molecules to be processed in batch. Accepted formats are plain text with one SMILES per line or SDF. For single molecules, users can also draw the molecule using the Kekule.js editor. (B) Results are summarized as a downloadable table with predictions of IC50 (μM) and whether the compound is likely to inhibit any of the 21 PPI complexes in this study. Other properties are also shown, and additional pharmacokinetic properties can be calculated using pkCSM.

achieved MCC values as high as 0.95 and 0.93 on training and nonredundant test sets, respectively, with the same set of features the distribution of IC50 values is the most diverse in the data set (Figures S6−S14) making the regression task significantly more difficult. Predictive models for inhibitors of the LEDGF/IN and HIF-1a/IN complexes achieved the lowest scores on the nonredundant blind tests, which is likely to be related to the small number of entries available for each target, 68 for the former and 82 for the latter. The best performing models with consistent predictions between 10-fold cross-validation and nonredundant blind tests are the Bcl2-Like/Bax-Bak, Mdm2-Like/P53, and Cyclophilins targets, the first two being well studied oncogenic PPIs with great implications to study of cancer therapies. The use of features for each target-specific regression model is summarized as Gini importance values output from the ExtraTrees algorithm in Tables S25−S33. As observed for the classification models previously described, overall importance for all regression models was spread across the different features.

Building a General PPI Inhibitor Predictor. Leveraging insights from the PPI-specific predictors, we developed a predictive model capable of identifying generic PPI inhibitors. For this purpose, we used all the available data from our original data set of inhibitors with experimental IC50 values, comprising 3972 targeting 45 distinct PPIs. After selecting features using our greedy stepwise approach (Table S34), performance was assessed by randomly selecting 80% of compounds for training the predictor and remaining 20% for testing, repeated 100 times (CV1). Here, our method achieved an average Pearson's correlation of 0.74 (σ = 0.02) and RMSE = 0.95 (σ = 0.07). On a similar configuration, but using a proportion of 50% of molecules in each set (CV2), we were able to achieve similar results with a Pearson's correlation and RMSE of 0.70 (σ = 0.01) and 1.07 (σ = 0.04), respectively, with small variations across 100 repetitions (Figure 3A).

Finally, we further evaluated our method on a low redundancy set in terms of molecule similarity. Here, we split our data set into 75% to train and remaining 25% as a test set, where all compounds in our test set have a similarity coefficient no greater than 0.80 (calculated via Tanimoto using Morgan fingerprints) to any molecules in the subset used for training the predictive model. This procedure resulted in a train set comprising 1581 small molecules and remaining 678 were used for validation. The distribution of IC50 values on each subset is depicted in Figure S15 in the Supporting Information. Under 10-fold cross-validation, our predictor was able to achieve Pearson's and Kendall's correlations of 0.75 and 0.56 (RMSE = 1.01), respectively, which is consistent with the results for the two cross-validation approaches discussed earlier (CV1 and CV2). After removing 10% of the outliers, the performance reached Pearson's and Kendall's correlations of 0.83 and 0.63 (RMSE = 0.86), respectively. On the low-redundant blind test set, our approach was able to achieve Pearson's and Kendall's correlations of 0.64 and 0.48 (RMSE = 1.08), respectively.

https://doi.org/10.1021/acs.jcim.1c01135
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

After removing 10% of the outliers, represented mostly by smaller compounds (average molecular weight of 425 Da), the predictive model achieved Pearson's and Kendall's correlations of 0.76 and 0.56 (RMSE = 0.74), respectively (Figure 3B). Finally, we evaluated the performance of our regressor in identifying more potent PPI inhibitors (IC50 < 1 μM) via classification by regression, and our predictive model showed a balanced prediction with an AUC of 0.77 and sensitivity and specificity of 76% and 78%, respectively. While not providing a definitive measurement of PPI affinity or inhibition constant, a robust general predictor can further our understanding of what makes a PPI inhibitor, also enabling complementary compound prioritization and ranking.

pdCSM-PPI Web Server. pdCSM-PPI is freely available via an easy-to-use web interface and API at http://biosig.unimelb.edu.au/pdcsm_ppi. Users can query the website with a single small molecule formatted as a SMILES string or provide a list of SMILES, one per line in a plain text file (Figure 4) for batch processing. For the single molecule submission, users have the option to draw the molecule using the Kekule.js editor.[30] On the results page, users have the option of calculating pharmacokinetic properties of selected molecules using pkCSM[19] and visualize molecule depiction via SmilesDrawer.[31] A full description with examples is available in the help page, and documentation for querying the web server using the API is available at http://biosig.unimelb.edu.au/pdcsm_ppi/docs.

CONCLUSIONS

Here, we present pdCSM-PPI, a novel predictive method to study the inhibitory activity of small molecules in PPI complexes. By implementing our well-established graph-based signatures framework with complementary physicochemical and fragment-based descriptors, we were able to outperform existing complex-specific methods in predicting inhibitors for nine different PPI targets: Bcl2-Like/Bak-Bax, Bromodomain/Histone, LEDGF/IN, LFA/ICAM, Mdm2-Like/P53, Ras/SOS1, WDR5/MLL, XIAP/Smac, and CD4/gp120. Overall, our approach can help identify PPI inhibitors for 21 different PPI complexes, representing the most comprehensive computational platform to assist PPI inhibitor development to date. Finally, we expanded our method into implementing a robust general predictor for PPI inhibitor potency that will complement compound prioritization. We anticipate pdCSM-PPI to be of great value in helping a more efficient and rapid screening of compounds targeting PPIs. The method is freely available as a web server, including an API for easy integration with other analysis pipelines, at http://biosig.unimelb.edu.au/pdcsm_ppi.

METHODS

Data Set. Experimentally characterized PPI inhibitors were retrieved from TIMBAL,[28] iPPI-DB,[16] and 2P2I-DB v2[15] (Figure S16), comprising 4965 small molecules targeting 51 distinct PPIs (Table S35). Inhibitors for each PPI were clustered separately using the Butina algorithm with a Tanimoto similarity cutoff of 0.8, based on Morgan molecular fingerprints (1024 bits) with a radius size of 2,[18] with one compound selected as a cluster representative to minimize redundancy. PPIs with at least 20 nonredundant inhibitors were selected to build class-specific predictors. In order to build more balanced and specific classifiers, for each PPI target, noninhibitors were selected in the same proportion as the number of inhibitors using the following approach: 50% of noninhibitors were randomly selected from compounds targeting single proteins on the PubChem database,[26] and the other half was randomly selected from the inhibitor pool. In order to minimize biased predictive models, data sets for 13 PPI targets with more than 40 nonredundant inhibitors were split based on ligand clustering,[32] where compounds are selected from larger to smaller clusters until 75% of entries in the data set were included in the training set, and the remaining 25% of the compounds in smaller clusters remained reserved as a blind test set. For 10 PPI targets where the number of nonredundant inhibitors ranged from 19 to 35, all the data were used for training the predictive models.

From the original data set, experimental half maximal inhibitory concentration (IC50) values were retrieved from TIMBAL and iPPI-DB for 3972 small molecules. After clustering (using the aforementioned procedure), a total of 2259 nonredundant small molecules were selected to build a general PPI inhibitor predictor. These were split into 75% for training the predictive model and remaining 25% as a holdout test set based on ligand clustering. All data used in this work are publicly available at http://biosig.unimelb.edu.au/pdcsm_ppi/data.

Graph-Based Signatures and Feature Engineering. We have previously described graph-based distance patterns[33,34] as a resourceful representation for modeling the structural environment and studying the effects of single and multiple point mutations on protein folding[35−41] and interactions.[42] Alternatively, these signatures have proven to be effective and powerful for the study of toxicity and pharmacokinetics[19−21] and more recently for identification of compounds likely to be active against specific Mycobacteria species.[22] In this study, each small molecule in the data set is modeled as an undirected, unweighted graph, where atoms are represented as nodes and labeled according to their properties as different pharmacophores (hydrogen donor, acceptor, hydrophobic, aromatic, positive ionizable, and negative ionizable). A distance matrix is calculated for each graph/compound via all-pairs shortest paths (Johnson's algorithm) using the implementation available on the iGraph library. The matrix is then used to extract cumulative distribution of distances between nodes (for different combinations of pharmacophore types) by varying a distance cutoff by 1 unit. Finally, cumulative distributions for all possible pairs of labels are compiled in a feature vector and used as evidence to train machine learning algorithms. Additional physicochemical properties and fragment-based descriptors calculated using the RDKit library for cheminformatics were included. Molecular substructure mining was implemented using the program MoSS.[43]

Model Selection and Validation. For regression and classification tasks presented in this study, we used the ExtraTrees algorithm available on the scikit-learn Python library with default parameters.[44] Model interpretability was investigated based on feature importance scores which are measured as the total reduction of the criterion brought by the feature (known as Gini importance). The performance for regression was evaluated using Pearson's, Spearman's, and Kendall's correlation coefficients, as well as root mean square error (RMSE). The classification task was carried out for each of the individual complexes in our data set, and performance was assessed based on F1-score, Matthew's correlation coefficient (MCC), and area under the receiving operator curve (AUC). Results are shown as 10-fold cross-validation on training and nonredundant blind tests at small molecule level. The effect of

https://doi.org/10.1021/acs.jcim.1c01135
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

outliers on regression was also investigated by assessing performance on 90% of the data.

Finally, for each predictive model, in order to minimize noise and identify an optimal combination of features, we performed an incremental stepwise greedy approach. For each feature in a signature vector, the performance is evaluated against the target value. The best performing feature is selected and "fixed". The process is repeated for each of the remaining features in combination with the previously selected in order to find the best pair of features. The procedure continues until all features are selected. For classification tasks, AUC was used as the metric to evaluate performance and Pearson's correlation coefficient on regression. In order to compare the number of features selected and performance of our bottom-up approach, here we evaluated the use of the recursive feature elimination (RFE)[45] algorithm using decision trees as the base learner for feature selection. For all predictive models (PPI-specific models and general regressor), the number of features selected using RFE was greatly superior to those selected with our greedy approach, and performance on the latter was overall better. Therefore, we opted to use less complex models (Occam's razor principle) using a smaller number of features selected via our approach.

Web Server. The server front end was developed using Materialize framework version 1.0.0, while the back end was built using Python via the Flask framework (version 1.0.2). The web server is hosted on a Linux server running Apache2.

DATA AND SOFTWARE AVAILABILITY

pdCSM-PPI predictive models were made freely available either as a user-friendly web interface and as an API for programmatic access at http://biosig.unimelb.edu.au/pdcsm_ppi. No login or license is required. All data sets used to train and validate predicted models are publicly available for download at http://biosig.unimelb.edu.au/pdcsm_ppi/data.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01135.

Description of data set used in this work, including distribution of inhibitors over different PPIs; performances of predictive models over 10-fold cross-validation and nonredundant test sets; description of features used for each predictive model, including feature importance (PDF)

AUTHOR INFORMATION

Corresponding Authors

Douglas E. V. Pires − *Systems and Computational Biology, Bio21 Institute, University of Melbourne, Parkville 3052 Victoria, Australia; Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne 3004 Victoria, Australia; School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane 4072, Australia; School of Computing and Information Systems, University of Melbourne, Parkville 3052 Victoria, Australia;* orcid.org/0000-0002-3004-2119; Email: douglas.pires@unimelb.edu.au

David B. Ascher − *Systems and Computational Biology, Bio21 Institute, University of Melbourne, Parkville 3052 Victoria, Australia; Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne 3004 Victoria, Australia; School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane 4072, Australia;* orcid.org/0000-0003-2948-2413; Phone: +61 90354794; Email: david.ascher@unimelb.edu.au

Author

Carlos H. M. Rodrigues − *Systems and Computational Biology, Bio21 Institute, University of Melbourne, Parkville 3052 Victoria, Australia; Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne 3004 Victoria, Australia; School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane 4072, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01135

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

C.H.M.R is funded by a Melbourne Research Scholarship. This work was supported part by the National Health and Medical Research Council of Australia (GNT1174405 to D.B.A.) and the Victorian Government's Operational Infrastructure Support Program.

REFERENCES

(1) Gao, J.; Li, W. X.; Feng, S. Q.; Yuan, Y. S.; Wan, D. F.; Han, W.; Yu, Y. A Protein-Protein Interaction Network of Transcription Factors Acting During Liver Cell Proliferation. *Genomics* **2008**, *91* (4), 347−355.

(2) Chuderland, D.; Seger, R. Protein-Protein Interactions in the Regulation of the Extracellular Signal-Regulated Kinase. *Mol. Biotechnol.* **2005**, *29* (1), 57−74.

(3) Paumi, C. M.; Menendez, J.; Arnoldo, A.; Engels, K.; Iyer, K. R.; Thaminy, S.; Georgiev, O.; Barral, Y.; Michaelis, S.; Stagljar, I. Mapping Protein-Protein Interactions for the Yeast Abc Transporter Ycf1p by Integrated Split-Ubiquitin Membrane Yeast Two-Hybrid Analysis. *Mol. Cell* **2007**, *26* (1), 15−25.

(4) Nicod, C.; Banaei-Esfahani, A.; Collins, B. C. Elucidation of Host-Pathogen Protein-Protein Interactions to Uncover Mechanisms of Host Cell Rewiring. *Curr. Opin. Microbiol.* **2017**, *39*, 7−15.

(5) Gao, M.; Zhou, H.; Skolnick, J. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure* **2015**, *23* (7), 1362−1369.

(6) David, A.; Razali, R.; Wass, M. N.; Sternberg, M. J. Protein-Protein Interaction Sites Are Hot Spots for Disease-Associated Nonsynonymous Snps. *Hum. Mutat.* **2012**, *33* (2), 359−363.

(7) Engin, H. B.; Kreisberg, J. F.; Carter, H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One* **2016**, *11* (4), No. e0152929.

(8) Jones, S.; Thornton, J. M. Prediction of Protein-Protein Interaction Sites Using Patch Analysis. *J. Mol. Biol.* **1997**, *272* (1), 133−143.

(9) Jubb, H.; Blundell, T. L.; Ascher, D. B. Flexibility and Small Pockets at Protein-Protein Interfaces: New Insights into Druggability. *Prog. Biophys. Mol. Biol.* **2015**, *119* (1), 2−9.

(10) Gandhi, L.; Camidge, D. R.; Ribeiro de Oliveira, M.; Bonomi, P.; Gandara, D.; Khaira, D.; Hann, C. L.; McKeegan, E. M.; Litvinovich, E.; Hemken, P. M.; Dive, C.; Enschede, S. H.; Nolan, C.; Chiu, Y. L.; Busman, T.; Xiong, H.; Krivoshik, A. P.; Humerickhouse, R.; Shapiro, G. I.; Rudin, C. M. Phase I Study of Navitoclax (Abt-263), a Novel Bcl-2 Family Inhibitor, in Patients with Small-Cell Lung Cancer and Other Solid Tumors. *J. Clin. Oncol.* **2011**, *29* (7), 909−916.

(11) Morelli, X.; Bourgeas, R.; Roche, P. Chemical and Structural Lessons from Recent Successes in Protein-Protein Interaction Inhibition (2p2i). *Curr. Opin. Chem. Biol.* **2011**, *15* (4), 475−481.
(12) Wilson, C. G.; Arkin, M. R. Small-Molecule Inhibitors of Il-2/Il-2r: Lessons Learned and Applied. *Curr. Top. Microbiol. Immunol.* **2010**, *348*, 25−59.
(13) Neugebauer, A.; Hartmann, R. W.; Klein, C. D. Prediction of Protein-Protein Interaction Inhibitors by Chemoinformatics and Machine Learning Methods. *J. Med. Chem.* **2007**, *50* (19), 4665−4668.
(14) Hamon, V.; Bourgeas, R.; Ducrot, P.; Theret, I.; Xuereb, L.; Basse, M. J.; Brunel, J. M.; Combes, S.; Morelli, X.; Roche, P. 2p2i Hunter: A Tool for Filtering Orthosteric Protein-Protein Interaction Modulators Via a Dedicated Support Vector Machine. *J. R. Soc., Interface* **2014**, *11* (90), 20130860.
(15) Basse, M. J.; Betzi, S.; Morelli, X.; Roche, P. 2p2idb V2: Update of a Structural Database Dedicated to Orthosteric Modulation of Protein-Protein Interactions. *Database* **2016**, *2016*, baw007.
(16) Torchet, R.; Druart, K.; Ruano, L.C.; Moine-Franel, A.; Borges, H.; Doppelt-Azeroual, O.; Brancotte, B.; Mareuil, F.; Nilges, M.; Menager, H.; Sperandio, O. The Ippi-Db Initiative: A Community-Centered Database of Protein-Protein Interaction Modulators. *Bioinformatics* **2021**, *37*, 89−96.
(17) Jana, T.; Ghosh, A.; Das Mandal, S.; Banerjee, R.; Saha, S. Ppimpred: A Web Server for High-Throughput Screening of Small Molecules Targeting Protein-Protein Interaction. *R. Soc. Open Sci.* **2017**, *4* (4), 160501.
(18) Gupta, P.; Mohanty, D. Smmppi: A Machine Learning-Based Approach for Prediction of Modulators of Protein-Protein Interactions and Its Application for Identification of Novel Inhibitors for Rbd:Hace2 Interactions in Sars-Cov-2. *Briefings Bioinf.* **2021**, *22*, bbab11.
(19) Pires, D. E.; Blundell, T. L.; Ascher, D. B. Pkcsm: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J. Med. Chem.* **2015**, *58* (9), 4066−4072.
(20) Kaminskas, L. M.; Pires, D. E. V.; Ascher, D. B. Dendpoint: A Web Resource for Dendrimer Pharmacokinetics Investigation and Prediction. *Sci. Rep.* **2019**, *9* (1), 15465.
(21) Pires, D. E. V.; Kaminskas, L. M.; Ascher, D. B. Prediction and Optimization of Pharmacokinetic and Toxicity Properties of the Ligand. *Methods Mol. Biol.* **2018**, *1762*, 271−284.
(22) Pires, D. E. V.; Ascher, D. B. Mycocsm: Using Graph-Based Signatures to Identify Safe Potent Hits against Mycobacteria. *J. Chem. Inf. Model.* **2020**, *60* (7), 3450−3456.
(23) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged Molecules for Protein Binding Identified from Nmr-Based Screening. *J. Med. Chem.* **2000**, *43* (18), 3443−3447.
(24) Whitty, A.; Kumaravel, G. Between a Rock and a Hard Place? *Nat. Chem. Biol.* **2006**, *2* (3), 112−118.
(25) Higueruelo, A. P.; Schreyer, A.; Bickerton, G. R.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Atomic Interactions and Profile of Small Molecules Disrupting Protein-Protein Interfaces: The Timbal Database. *Chem. Biol. Drug Des.* **2009**, *74* (5), 457−467.
(26) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. Pubchem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49* (D1), D1388−D1395.
(27) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. Chembl: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930−D940.
(28) Higueruelo, A. P.; Jubb, H.; Blundell, T. L. Timbal V2: Update of a Database Holding Small Molecules Modulating Protein-Protein Interactions. *Database* **2013**, *2013*, No. bat039.
(29) Kuenemann, M. A.; Bourbon, L. M.; Labbe, C. M.; Villoutreix, B. O.; Sperandio, O. Which Three-Dimensional Characteristics Make Efficient Inhibitors of Protein-Protein Interactions? *J. Chem. Inf. Model.* **2014**, *54* (11), 3067−3079.
(30) Jiang, C.; Jin, X.; Dong, Y.; Chen, M. Kekule. Js: An Open Source Javascript Chemoinformatics Toolkit. *J. Chem. Inf. Model.* **2016**, *56* (6), 1132−1138.
(31) Probst, D.; Reymond, J. L. Smilesdrawer: Parsing and Drawing Smiles-Encoded Molecular Structures Using Client-Side Javascript. *J. Chem. Inf. Model.* **2018**, *58* (1), 1−7.
(32) Martin, E. J.; Polyakov, V. R.; Zhu, X. W.; Tian, L.; Mukherjee, P.; Liu, X. All-Assay-Max2 Pqsar: Activity Predictions as Accurate as Four-Concentration Ic50s for 8558 Novartis Assays. *J. Chem. Inf. Model.* **2019**, *59* (10), 4450−4459.
(33) Pires, D. E.; de Melo-Minardi, R. C.; dos Santos, M. A.; da Silveira, C. H.; Santoro, M. M.; Meira, W., Jr Cutoff Scanning Matrix (Csm): Structural Classification and Function Prediction by Protein Inter-Residue Distance Patterns. *BMC Genomics* **2011**, *12* (S4), S12.
(34) Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Campos, F. F.; Meira, W., Jr Acsm: Noise-Free Graph-Based Signatures to Large-Scale Receptor-Based Ligand Prediction. *Bioinformatics* **2013**, *29* (7), 855−861.
(35) Pires, D. E.; Ascher, D. B.; Blundell, T. L. Mcsm: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* **2014**, *30* (3), 335−342.
(36) Pires, D. E.; Ascher, D. B.; Blundell, T. L. Duet: A Server for Predicting Effects of Mutations on Protein Stability Using an Integrated Computational Approach. *Nucleic Acids Res.* **2014**, *42* (W1), W314−319.
(37) Rodrigues, C. H.; Pires, D. E.; Ascher, D. B. Dynamut: Predicting the Impact of Mutations on Protein Conformation, Flexibility and Stability. *Nucleic Acids Res.* **2018**, *46* (W1), W350−W355.
(38) Rodrigues, C. H. M.; Pires, D. E. V.; Ascher, D. B. Dynamut2: Assessing Changes in Stability and Flexibility Upon Single and Multiple Point Missense Mutations. *Protein Sci.* **2021**, *30* (1), 60−69.
(39) Pires, D. E.; Ascher, D. B. Mcsm-Ab: A Web Server for Predicting Antibody-Antigen Affinity Changes Upon Mutation with Graph-Based Signatures. *Nucleic Acids Res.* **2016**, *44* (W1), W469−473.
(40) Myung, Y.; Rodrigues, C. H. M.; Ascher, D. B.; Pires, D. E. V. mCSM-Ab2: Guiding Rational Antibody Design Using Graph-Based Signatures. *Bioinformatics* **2020**, *36* (5), 1453−1459.
(41) Myung, Y.; Pires, D. E. V.; Ascher, D. B. Mmcsm-Ab: Guiding Rational Antibody Engineering through Multiple Point Mutations. *Nucleic Acids Res.* **2020**, *48* (W1), W125−W131.
(42) Rodrigues, C. H. M.; Myung, Y.; Pires, D. E. V.; Ascher, D. B. Mcsm-Ppi2: Predicting the Effects of Mutations on Protein-Protein Interactions. *Nucleic Acids Res.* **2019**, *47* (W1), W338−W344.
(43) Borgelt, C., Meinl, T., Berthold, M. Moss: A Program for Molecular Substructure Mining. In *1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, **2005**; p 6.
(44) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.
(45) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine learning* **2002**, *46*, 389−422.

https://doi.org/10.1021/acs.jcim.1c01135
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

SUPPORTING INFORMATION

pdCSM-PPI: Using Graph-Based Signatures to Identify Protein-Protein Interaction Inhibitors

Carlos H. M. Rodrigues [1,2,3], Douglas E. V. Pires [1,2,3,4*], David B. Ascher [1,2,3*]

[1] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Parkville 3052, Victoria, Australia

[2] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne 3004, Victoria, Australia

[3] School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, 4072 Australia

[4] School of Computing and Information Systems, University of Melbourne, Parkville 3052, Victoria, Australia

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

TABLES

Table S1. Feature importance for predictive model of PPI inhibitors targeting the Bcl2/Bak complex.

Feature	Description	Importance
Aromatic_Count	Number of atoms in aromatic rings	0.105
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe atoms within 4 bonds	0.055
Chi2n	Molecular connectivity index	0.052
fr_sulfide	Number of thioether	0.051
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.046
fr_Ar_N	Number of aromatic nitrogens	0.042
fr_bicyclic	Number of bicyclic structures	0.042
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.037
Acceptor:Hydrophobe-3.00	Number of pairs of acceptor-hydrophobe atoms within 3 bonds	0.028
VSA_EState8	MOE-type descriptors, surface area contributions and EState indices	0.027
SlogP_VSA12	MOE-type descriptors, LogP and surface area contributions	0.026
SMR_VSA9	MOE-type descriptors, LogP and surface area contributions	0.025
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.022
Donor:Donor-3.00	Number of pairs of hydrophobe atoms within 3 bonds	0.022

fr_thiazole	Number of thiazole rings	0.021
Donor_Count	Number hydrogen bond donors	0.020
SMR_VSA5	MOE-type descriptors, molar refractivity and surface area contributions	0.019
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.019
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.018
Donor:Hydrophobe-3.00	Number of pairs of donor-hydrophobe atoms within 3 bond	0.018
fr_pyridine	Number of pyridine rings	0.017
fr_phenol	Number of phenolic OH excluding ortho intramolecular Hbond substituents	0.017
PEOE_VSA9	MOE-type descriptors, partial charges and surface area contributions	0.016
SlogP_VSA7	MOE-type descriptors, LogP and surface area contributions	0.016
Donor:Donor-6.00	Number of pairs of hydrogen donor within 6 bonds	0.016
PosIonizable_Count	Number of PosIonizable atoms	0.015
fr_Ar_OH	Number of aromatic hydroxyl groups	0.014
Donor:Donor-1.00	Number of pairs of hydrogen donor atoms within 1 bond	0.012
Aromatic:NegIonizable-4.00	Number of pairs of aromatic-negIonizable atoms within 4 bonds	0.011
fr_Imine	Number of Imines	0.009
fr_NH2	Number of Primary amines	0.009
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-negIonizable atoms within 4 bonds	0.008
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonizable atoms within 1 bond	0.008
Acceptor:PosIonizable-4.00	Number of pairs of acceptor-posIonizable atoms within 4 bonds	0.008
NegIonizable:NegIonizable-1.00	Number of pairs of negIonizable-negIonizable atoms within 1 bond	0.008
PosIonizable:PosIonizable-4.00	Number of pairs of posIonizable-posIonizable atoms within 4 bonds	0.007
fr_alkyl_halide	Number of alkyl halides	0.007
NegIonizable:NegIonizable-4.00	Number of pairs of negIonizable-negIonizable atoms within 4 bonds	0.007
fr_unbrch_alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)	0.007
NegIonizable:NegIonizable-2.00	Number of pairs of negIonizable-negIonizable atoms within 2 bonds	0.007
NegIonizable:NegIonizable-5.00	Number of pairs of negIonizable-negIonizable atoms within 5 bonds	0.006
Donor:PosIonizable-3.00	Number of pairs of donor-posIonizable atoms within 3 bonds	0.006

Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonizable atoms within 3 bonds	0.006
PosIonizable:PosIonizable-3.00	Number of pairs of posIonizable-posIonizable atoms within 3 bonds	0.006
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonizable atoms within 1 bond	0.005
Hydrophobe:NegIonizable-4.00	Number of pairs of hydrophobe-negIonizable atoms within 4 bonds	0.005
fr_ketone	Number of ketones	0.005
PosIonizable:PosIonizable-2.00	Number of pairs of posIonizable-posIonizable atoms within 2 bonds	0.005
Hydrophobe:NegIonizable-2.00	Number of pairs of hydrophobe-negIonizable atoms within 2 bonds	0.005
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonizable atoms within 1 bond	0.004
fr_priamide	Number of primary amides	0.004
Donor:PosIonizable-1.00	Number of pairs of donor-posIonizable atoms within 1 bond	0.004
Acceptor:NegIonizable-4.00	Number of pairs of acceptor-negIonizable atoms within 4 bonds	0.004
Acceptor:NegIonizable-2.00	Number of pairs of acceptor-negIonizable atoms within 2 bonds	0.004
Donor:NegIonizable-5.00	Number of pairs of acceptor-negIonizable atoms within 5 bonds	0.004
Tox_2	[CH]=[CH]O	0.004
fr_thiophene	Number of thiophene rings	0.003
Donor:NegIonizable-1.00	Number of pairs of donor-negIonizable atoms within 1 bond	0.003
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonizable atoms within 1 bond	0.003
Donor:NegIonizable-3.00	Number of pairs of donor-posIonizable atoms within 3 bonds	0.003
fr_guanido	Number of guanidine groups	0.001
fr_lactone	Number of cyclic esters (lactones)	0.001

Table S2. Feature importance for predictive model of PPI inhibitors targeting the Bromodomain/Histone complex.

Feature	**Description**	**Importance**
Donor:Hydrophobe-4.00	Number of pairs of donor-hydrophobe atoms within 4 bonds	0.044
Kappa2	Molecular shape index	0.043
fr_aryl_methyl	Number of aryl methyl sites for hydroxylation	0.041
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.037
BalabanJ	Balaban's connectivity topological index	0.033
fr_Al_COO	Number of aliphatic carboxylic acids	0.024

fr_halogen	Number of halogens	0.024
VSA_EState10	MOE-type descriptors surface area contributions and EState indices	0.024
Chi1v	Molecular connectivity index	0.023
SMR_VSA5	MOE-type descriptors, molar refractivity and surface area contributions	0.022
fr_aniline	Number of anilines	0.022
Aromatic:Aromatic-5.00	Number of pairs of aromatic-aromatic atoms within 5 bonds	0.021
Aromatic:Donor-3.00	Number of pairs of aromatic-donor atoms within 3 bonds	0.021
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophobe atoms within 4 bonds	0.020
fr_benzene	Number of benzene rings	0.020
VSA_EState8	MOE-type descriptors surface area contributions and EState indices	0.018
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.018
PEOE_VSA2	MOE-type descriptors, partial charges and surface area contributions	0.017
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.017
Acceptor:Aromatic-5.00	Number of pairs of acceptor-aromatic atoms within 5 bonds	0.017
SMR_VSA6	MOE-type descriptors, molar refractivity and surface area contributions	0.017
Aromatic:Hydrophobe-5.00	Number of pairs of aromatic-hydrophobe atoms within 5 bonds	0.017
Aromatic_Count	Number of atoms in aromatic rings	0.016
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.016
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.016
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.015
fr_NH0	Number of Tertiary amines	0.015
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.015
SMR_VSA4	MOE-type descriptors, molar refractivity and surface area contributions	0.015
Acceptor:Aromatic-4.00	Number of pairs of acceptor-aromatic atoms within 4 bonds	0.015
PEOE_VSA7	MOE-type descriptors, partial charges and surface area contributions	0.014
Acceptor:Aromatic-2.00	Number of pairs of acceptor-aromatic atoms within 2 bonds	0.014
Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.014

PEOE_VSA3	MOE-type descriptors, partial charges and surface area contributions	0.014
PEOE_VSA9	MOE-type descriptors, partial charges and surface area contributions	0.013
SlogP_VSA1	MOE-type descriptors, LogP and surface area contributions	0.013
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.013
Aromatic:Donor-1.00	Number of pairs of aromatic-donor atoms within 1 bond	0.013
PEOE_VSA11	MOE-type descriptors, partial charges and surface area contributions	0.012
PEOE_VSA4	MOE-type descriptors, partial charges and surface area contributions	0.012
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.012
Acceptor:Acceptor-5.00	Number of pairs of acceptor-acceptor atoms within 5 bonds	0.012
NegIonizable_Count	Number of negIonizable atoms	0.011
PosIonizable_Count	Number of posIonizable atoms	0.010
Donor:Donor-4.00	Number of pairs of donor-donor atoms within 4 bonds	0.010
PEOE_VSA5	MOE-type descriptors, partial charges and surface area contributions	0.009
Aromatic:PosIonizable-6.00	Number of pairs of aromatic-posIonizable atoms within 6 bonds	0.008
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.008
fr_Imine	Number of Imines	0.007
fr_Ar_OH	Number of aromatic hydroxyl groups	0.007
fr_ketone	Number of ketones	0.007
fr_thiophene	Number of thiophene rings	0.006
Acceptor:PosIonizable-6.00	Number of pairs of acceptor-posIonazable atoms within 6 bonds	0.006
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.006
Aromatic:PosIonizable-5.00	Number of pairs of acceptor-posIonazable atoms within 5 bonds	0.006
Aromatic:PosIonizable-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.006
fr_unbrch_alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)	0.005
fr_nitrile	Number of nitriles	0.005
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonazable atoms within 1 bond	0.005
fr_nitro_arom	Number of nitro benzene ring substituents	0.004
Tox_2	[CH]=[CH]O	0.004
fr_nitro	Number of nitro groups	0.004

Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.004
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.004
fr_azo	Number of azo groups	0.004
Donor:PosIonizable-4.00	Number of pairs of donor-negIonazable atoms within 2 bonds	0.004
PosIonizable:PosIonizable-5.00	Number of pairs of posIonazable-posIonazable atoms within 5 bonds	0.004
Acceptor:PosIonizable-3.00	Number of pairs of acceptor-posIonazable atoms within 3 bonds	0.004
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.003
Acceptor:PosIonizable-2.00	Number of pairs of acceptor-posIonazable atoms within 2 bonds	0.003
fr_hdrzone	Number of hydrazone groups	0.003
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.003
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.002
Aromatic:NegIonizable-1.00	Number of pairs of aromatic-negIonazable atoms within 1 bond	0.002
fr_C_S	Number of thiocarbonyl	0.002
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.002
Donor:NegIonizable-2.00	Number of pairs of donor-negIonazable atoms within 2 bonds	0.002

Table S3. Feature importance for predictive model of PPI inhibitors targeting the CD4/gp120 complex.

Feature	**Description**	**Importance**
fr_piperzine	Number of piperzine rings	0.559
PEOE_VSA13	MOE-type descriptors, partial charges and surface area contributions	0.270
fr_amide	Number of amides	0.169
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.002

Table S4. Feature importance for predictive model of PPI inhibitors targeting the Cyclophilins complex.

Feature	**Description**	**Importance**
fr_amide	Number of amides	0.102
Aromatic:Donor-1.00	Number of pairs of aromatic-donor atoms within 1 bond	0.074
fr_Ar_N	Number of aromatic nitrogens	0.067
Chi1n	Molecular connectivity index	0.059
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.054

Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.046
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.044
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.039
fr_benzene	Number of benzene rings	0.038
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.033
fr_pyridine	Number of pyridine rings	0.029
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.027
PosIonizable:PosIonizable-2.00	Number of pairs of posIonizable-posIonizable atoms within 2 bonds	0.027
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.027
Aromatic:Hydrophobe-2.00	Number of pairs of aromatic-hydrophobe atoms within 2 bonds	0.026
SMR_VSA6	MOE-type descriptors, molar refractivity and surface area contributions	0.025
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.025
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.023
PosIonizable:PosIonizable-1.00	Number of pairs of posIonizable-posIonizable atoms within 1 bond	0.021
Acceptor:Acceptor-4.00	Number of pairs of acceptor-acceptor atoms within 4 bonds	0.021
Hydrophobe:Hydrophobe-2.00	Number of pairs of hydrophobe-hydrophobe atoms within 2 bonds	0.019
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.019
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonizable atoms within 1 bond	0.017
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.017
Acceptor:Acceptor-2.00	Number of pairs of acceptor-acceptor atoms within 2 bonds	0.016
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.015
Donor:PosIonizable-1.00	Number of pairs of donor-posIonizable atoms within 1 bond	0.014
Hydrophobe:NegIonizable-2.00	Number of pairs of hydrophobe-negIonizable atoms within 2 bonds	0.012
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonizable atoms within 1 bond	0.011
Donor:PosIonizable-2.00	Number of pairs of donor-posIonizable atoms within 2 bonds	0.011
NegIonizable:NegIonizable-1.00	Number of pairs of negIonizable-negIonizable	0.008

	atoms within 1 bond	
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonizable atoms within 1 bond	0.006
Donor:NegIonizable-2.00	Number of pairs of donor-negIonizable atoms within 2 bonds	0.006
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonizable atoms within 2 bonds	0.005
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonizable atoms within 1 bond	0.005
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonizable atoms within 1 bond	0.003
Donor:NegIonizable-3.00	Number of pairs of donor-negIonizable atoms within 3 bonds	0.003
Donor:NegIonizable-1.00	Number of pairs of donor-negIonizable atoms within 1 bond	0.002
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonizable atoms within 2 bonds	0.002
Aromatic:NegIonizable-1.00	Number of pairs of aromatic-negIonizable atoms within 1 bond	0.001

Table S5. Feature importance for predictive model of PPI inhibitors targeting the Fkbp1a/Fk506 complex.

Feature	**Description**	**Importance**
NHOHCount	Number of NHs or OHs	0.103
Acceptor:Acceptor-3.00	Number of pairs of acceptor-acceptor atoms within 3 bonds	0.093
PosIonizable_Count	Number of posIonizable atoms	0.092
fr_bicyclic	Number of bicyclic structure	0.081
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.072
Aromatic_Count	Number of atoms in aromatic rings	0.059
fr_methoxy	Number of methoxy groups -OCH3	0.059
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.053
fr_NH0	Number of Tertiary amines	0.051
Aromatic:Hydrophobe-4.00	Number of pairs of aromatic-hydrophobe atoms within 4 bonds	0.050
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.032
Donor:Hydrophobe-5.00	Number of pairs of donor-hydrophobe atoms within 5 bonds	0.029
Hydrophobe:Hydrophobe-2.00	Number of pairs of hydrophobe-hydrophobe atoms within 2 bonds	0.028
Acceptor:Aromatic-1.00	Number of pairs of acceptor-aromatic atoms within 1 bond	0.024
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.023

Donor:Hydrophobe-2.00	Number of pairs of donor-hydrophobe atoms within 2 bonds	0.017
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.016
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.013
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.012
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.010
Aromatic:PosIonizable-2.00	Number of pairs of aromatic-posIonazable atoms within 2 bonds	0.010
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.008
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.007
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.006
NegIonizable:NegIonizable-3.00	Number of pairs of negIonazable-negIonazable atoms within 3 bonds	0.006
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.005
fr_oxazole	Number of oxazole rings	0.005
NegIonizable:NegIonizable-4.00	Number of pairs of negIonazable-negIonazable atoms within 4 bonds	0.005
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.004
fr_sulfone	Number of sulfone groups	0.004
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonazable atoms within 1 bond	0.004
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.003
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.003
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.003
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.002
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.002
Aromatic:NegIonizable-3.00	Number of pairs of aromatic-negIonazable atoms within 3 bonds	0.002
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.001
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.001

Table S6. Feature importance for predictive model of PPI inhibitors targeting the HIF1-alpha/p300 complex.

Feature	Descriptor	Importance
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.234
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.127
RingCount	Number of aromatic rings	0.126
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.100
VSA_EState10	MOE-type descriptors, surface area contributions and EState indices	0.081
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.061
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.049
fr_Al_OH	Number of aliphatic hydroxyl groups	0.046
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.032
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.029
Donor:PosIonizable-6.00	Number of pairs of donor-posIonazable atoms within 6 bonds	0.027
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.021
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.018
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.016
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.013
Aromatic:NegIonizable-1.00	Number of pairs of aromatic-negIonazable atoms within 1 bond	0.010
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.005
fr_alkyl_carbamate	Number of alkyl carbamates (subject to hydrolysis)	0.002
NegIonizable:PosIonizable-1.00	Number of pairs of negIonazable-posIonazable atoms within 1 bond	0.001
NegIonizable:PosIonizable-2.00	Number of pairs of negIonazable-posIonazable atoms within 1 bonds	0.001

Table S7. Feature importance for predictive model of PPI inhibitors targeting the Integrins complex.

Feature	Description	Importance
fr_Al_COO	Number of aliphatic carboxylic acids	0.245
PEOE_VSA14	MOE-type descriptors, partial charges and surface area contributions	0.050

Donor:NegIonizable-2.00	Number of pairs of donor-negIonazable atoms within 2 bonds	0.039
fr_NH1	Number of Secondary amines	0.036
PEOE_VSA12	MOE-type descriptors, partial charges and surface area contributions	0.035
Hydrophobe:NegIonizable-4.00	Number of pairs of hydrophobe-negIonazable atoms within 4 bonds	0.035
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.033
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.031
PEOE_VSA13	MOE-type descriptors, partial charges and surface area contributions	0.031
Hydrophobe:NegIonizable-2.00	Number of pairs of hydrophobe-negIonazable atoms within 2 bonds	0.029
Aromatic_Count	Number of atoms in aromatic rings	0.024
SlogP_VSA2	MOE-type descriptors, LogP and surface area contributions	0.024
MolWt	Molecular weight	0.023
PEOE_VSA2	MOE-type descriptors, partial charges and surface area contributions	0.022
fr_amidine	Number of amidine groups	0.021
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.020
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophobe atoms within 4 bonds	0.018
HallKierAlpha	Hall-Kier alpha value	0.016
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.014
Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.014
fr_pyridine	Number of pyridine rings	0.014
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.013
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.013
Acceptor:Aromatic-5.00	Number of pairs of acceptor-aromatic atoms within 5 bonds	0.013
SlogP_VSA4	MOE-type descriptors, LogP and surface area contributions	0.012
Acceptor:Acceptor-6.00	Number of pairs of acceptor-acceptor atoms within 6 bonds	0.012
Donor:Donor-4.00	Number of pairs of donor-donor atoms within 4 bonds	0.011
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.011
FCount	Number of fluorine atoms	0.010

fr_barbitur	Number of barbiturate groups	0.007
fr_alkyl_halide	Number of alkyl halides	0.007
fr_Imine	Number of Imines	0.007
fr_sulfonamd	Number of sulfonamides	0.007
fr_piperdine	Number of piperdine rings	0.007
fr_imide	Number of imide groups	0.007
Aromatic:PosIonizable-6.00	Number of pairs of aromatic-posIonazable atoms within 6 bonds	0.006
fr_ArN	Number of N functional groups attached to aromatics	0.006
Donor:PosIonizable-4.00	Number of pairs of donor-posIonazable atoms within 4 bonds	0.005
fr_Ndealkylation1	Number of XCCNR groups	0.005
Donor:PosIonizable-6.00	Number of pairs of donor-posIonazable atoms within 6 bonds	0.005
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.005
Donor:PosIonizable-5.00	Number of pairs of donor-posIonazable atoms within 5 bonds	0.005
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.005
Acceptor:PosIonizable-6.00	Number of pairs of acceptor-posIonazable atoms within 6 bonds	0.005
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.005
Hydrophobe:PosIonizable-5.00	Number of pairs of hydrophobe-posIonazable atoms within 5 bonds	0.005
Aromatic:PosIonizable-2.00	Number of pairs of aromatic-posIonazable atoms within 2 bonds	0.004
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazble-posIonazable atoms within 3 bonds	0.004
Acceptor:PosIonizable-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.004
PosIonizable:PosIonizable-5.00	Number of pairs of posIonazble-posIonazable atoms within 5 bonds	0.004
PosIonizable:PosIonizable-6.00	Number of pairs of posIonazble-posIonazable atoms within 6 bonds	0.004
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.003
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.002
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.002
fr_nitrile	Number of nitriles	0.002
Tox_31	1~a~a~a~a2~a~1~a3~a(~a~2)~a~a~a~a~3	0.001
fr_aldehyde	Number of aldehydes	0.001
NegIonizable:PosIonizable-2.00	Number of pairs of negIonazable-	0.001

	posIonazable atoms within 2 bonds	
fr_SH	Number of thiol groups	0.001
fr_lactone	Number of cyclic esters (lactones)	0.001
NegIonizable:PosIonizable-6.00	Number of pairs of negIonazable-posIonazable atoms within 6 bonds	0.001

Table S8. Feature importance for predictive model of PPI inhibitors targeting the Ledgf/IN complex.

Feature	**Description**	**Importance**
Acceptor:Aromatic-4.00	Number of pairs of acceptor-aromatic atoms within 4 bonds	0.059
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.052
fr_C_O_noCOO	Number of carbonyl O, excluding COOH	0.045
fr_amide	Number of amides	0.038
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.037
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.035
Acceptor:Hydrophobe-4.00	Number of pairs of acceptor-hydrophobe atoms within 4 bonds	0.035
fr_pyridine	Number of pyridine rings	0.032
Acceptor:Acceptor-5.00	Number of pairs of acceptor-acceptor atoms within 5 bonds	0.029
PEOE_VSA2	MOE-type descriptors, partial charges and surface area contributions	0.024
Acceptor:Hydrophobe-5.00	Number of pairs of acceptor-hydrophobe atoms within 5 bonds	0.024
HallKierAlpha	Hall-Kier alpha value	0.022
Hydrophobe:Hydrophobe-3.00	Number of pairs of hydrophobe-hydrophobe atoms within 3 bonds	0.022
PEOE_VSA10	MOE-type descriptors, partial charges and surface area contributions	0.022
Chi2n	Molecular connectivity index	0.021
Donor:NegIonizable-4.00	Number of pairs of donor-negIonazable atoms within 4 bonds	0.021
SlogP_VSA2	MOE-type descriptors, LogP and surface area contributions	0.020
SMR_VSA2	MOE-type descriptors, molar refractivity and surface area contributions	0.019
SlogP_VSA4	MOE-type descriptors, LogP and surface area contributions	0.019
Chi0v	Molecular connectivity index	0.019
PEOE_VSA6	MOE-type descriptors, partial charges and surface area contributions	0.019
Chi0	Molecular connectivity index	0.018
Kappa1	Molecular shape index	0.018

Aromatic:Hydrophobe-4.00	Number of pairs of aromatic-hydrophobe atoms within 4 bonds	0.017
Chi1v	Molecular connectivity index	0.017
PEOE_VSA7	MOE-type descriptors, partial charges and surface area contributions	0.015
Donor:Hydrophobe-5.00	Number of pairs of donor-hydrophobe atoms within 5 bonds	0.015
PEOE_VSA12	MOE-type descriptors, partial charges and surface area contributions	0.015
Aromatic:Donor-3.00	Number of pairs of aromatic-donor atoms within 3 bonds	0.015
fr_nitrile	Number of nitriles	0.013
Acceptor:Donor-5.00	Number of pairs of acceptor-donor atoms within 5 bonds	0.013
Acceptor:Donor-2.00	Number of pairs of acceptor-donor atoms within 2 bonds	0.013
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.012
fr_Ndealkylation2	Number of tert-alicyclic amines (no heteroatoms, not quinine-like bridged N)	0.011
Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.011
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.011
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.010
fr_C_S	Number of thiocarbonyl	0.009
fr_hdrzone	Number of hydrazone groups	0.009
fr_urea	Number of urea groups	0.009
fr_quatN	Number of quarternary nitrogens	0.008
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.008
PEOE_VSA5	MOE-type descriptors, partial charges and surface area contributions	0.008
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.008
SlogP_VSA7	MOE-type descriptors, LogP and surface area contributions	0.008
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.007
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.006
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.006
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.006
fr_ArN	Number of N functional groups attached to aromatics	0.005

fr_unbrch_alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)	0.005
fr_priamide	Number of primary amides	0.005
fr_alkyl_halide	Number of alkyl halides	0.005
fr_nitro_arom	Number of nitro benzene ring substituents	0.005
fr_nitro	Number of nitro groups	0.005
fr_ketone	Number of ketones	0.005
Tox_1	O=N(~O)a	0.005
Tox_2	a[NH2]	0.004
fr_Imine	Number of Imines	0.004
fr_Al_OH_noTert	Number of aliphatic hydroxyl groups excluding tert-OH	0.003
Donor:NegIonizable-3.00	Number of pairs of donor-negIonazable atoms within 3 bonds	0.003
Tox_12	[OH,NH2][N,O]	0.002
fr_azo	Number of azo groups	0.002
fr_Al_OH	Number of aliphatic hydroxyl groups	0.002
fr_oxime	Number of oxime groups	0.002
fr_thiazole	Number of thiazole rings	0.002
fr_nitro_arom_nonortho	Number of non-ortho nitro benzene ring substituents	0.001
fr_Ndealkylation1	Number of XCCNR groups	0.001

Table S9. Feature importance for predictive model of PPI inhibitors targeting the Lfa/Icam complex.

Feature	**Description**	**Importance**
VSA_EState8	MOE-type descriptors, surface area contributions and EState indices	0.132
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.121
fr_amide	Number of amides	0.115
SMR_VSA5	MOE-type descriptors, molar refractivity and surface area contributions	0.089
Kappa2	Molecular shape index	0.081
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.074
fr_bicyclic	Number of bicyclic structures	0.073
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.072
fr_imide	Number of imide groups	0.065
Acceptor:Donor-4.00	Number of pairs of acceptor-donor atoms within 4 bonds	0.063
fr_para_hydroxylation	Number of para-hydroxylation sites	0.045
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.034
fr_urea	Number of urea groups	0.026

fr_guanido	Number of guanidine groups	0.009
fr_term_acetylene	Number of terminal acetylenes	0.002

Table S10. Feature importance for predictive model of PPI inhibitors targeting the Mdm2-like/P53 complex.

Feature	Description	Importance
VSA_EState10	MOE-type descriptors, surface area contributions and EState indices	0.085
fr_halogen	Number of halogens	0.051
SMR_VSA7	MOE-type descriptors, molar refractivity and surface area contributions	0.045
Donor:Hydrophobe-4.00	Number of pairs of donor-hydrophobe atoms within 4 bonds	0.040
Aromatic:Aromatic-6.00	Number of pairs of aromatic-aromatic atoms within 6 bonds	0.031
Aromatic:Aromatic-5.00	Number of pairs of aromatic-aromatic atoms within 6 bonds	0.026
Chi4n	Molecular connectivity index	0.025
SMR_VSA9	MOE-type descriptors, molar refractivity and surface area contributions	0.023
Aromatic_Count	Number of atoms in aromatic rings	0.022
Aromatic:Aromatic-4.00	Number of pairs of aromatic-aromatic atoms within 4 bonds	0.021
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.020
fr_aniline	Number of anilines	0.020
PEOE_VSA8	MOE-type descriptors, partial charges and surface area contributions	0.018
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.018
Chi3n	Molecular connectivity index	0.018
MolWt	Molecular weight	0.017
fr_NH1	Number of Secondary amines	0.017
BalabanJ	Balaban's connectivity topological index	0.017
PEOE_VSA13	MOE-type descriptors, partial charges and surface area contributions	0.017
fr_bicyclic	Number of bicyclic structures	0.016
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.015
Acceptor:Aromatic-1.00	Number of pairs of acceptor-aromatic atoms within 1 bond	0.015
Chi0n	Molecular connectivity index	0.014
HeavyAtomCount	Number of heavy atoms	0.014
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.014
fr_aryl_methyl	Number of aryl methyl sites for hydroxylation	0.013

fr_C_O_noCOO	Number of carbonyl O, excluding COOH	0.012
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.012
Acceptor:Acceptor-3.00	Number of pairs of acceptor-acceptor atoms within 3 bonds	0.012
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-posIonazable atoms within 4 bonds	0.012
Acceptor:Aromatic-3.00	Number of pairs of acceptor-aromatic atoms within 3 bonds	0.012
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.011
SMR_VSA1	MOE-type descriptors, molar refractivity and surface area contributions	0.011
Acceptor:Hydrophobe-6.00	Number of pairs of acceptor-hydrophobe atoms within 6 bonds	0.011
VSA_EState8	MOE-type descriptors surface area contributions and EState indices	0.011
PEOE_VSA14	MOE-type descriptors, partial charges and surface area contributions	0.011
fr_C_O	Number of carbonyl O	0.011
Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.011
Kappa1	Molecular shape index	0.010
fr_pyridine	Number of pyridine rings	0.010
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.010
PEOE_VSA11	MOE-type descriptors, partial charges and surface area contributions	0.010
Acceptor:Hydrophobe-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.009
fr_morpholine	Number of morpholine rings	0.009
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.009
VSA_EState9	MOE-type descriptors surface area contributions and EState indices	0.009
Donor_Count	Number of hydrogen donors	0.008
PEOE_VSA9	MOE-type descriptors, partial charges and surface area contributions	0.008
Aromatic:PosIonizable-3.00	Number of pairs of aromatic-posIonazable atoms within 3 bonds	0.008
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.008
fr_piperzine	Number of piperzine rings	0.007
Donor:Donor-4.00	Number of pairs of donor-donor atoms within 4 bonds	0.007
Acceptor:PosIonizable-5.00	Number of pairs of acceptor-posIonazable atoms within 5 bonds	0.007
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3	0.006

	bonds	
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.006
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.006
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.005
fr_Ndealkylation1	Number of XCCNR groups	0.005
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.005
fr_phenol	Number of phenols	0.005
FCount	Number of fluorine atoms	0.005
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.004
NegIonizable:NegIonizable-6.00	Number of pairs of negIonazable-negIonazable atoms within 6 bonds	0.004
Donor:PosIonizable-5.00	Number of pairs of donor-negIonazable atoms within 5 bonds	0.004
NegIonizable:NegIonizable-3.00	Number of pairs of negIonazable-negIonazable atoms within 3 bonds	0.004
fr_nitrile	Number of nitriles	0.004
fr_NH2	Number of Primary amines	0.003
Donor:PosIonizable-3.00	Number of pairs of donor-posIonazable atoms within 3 bonds	0.003
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonazable atoms within 1 bond	0.003
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.003
Tox_1	O=N(~O)a	0.003
Donor:NegIonizable-2.00	Number of pairs of donor-negIonazable atoms within 2 bonds	0.003
fr_imide	Number of imide groups	0.002
fr_nitro	Number of nitro groups	0.002
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.002
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.002
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonazable atoms within 2 bonds	0.002
fr_nitro_arom	Number of nitro benzene ring substituents	0.002
Aromatic:NegIonizable-4.00	Number of pairs of aromatic-negIonazable atoms within 4 bonds	0.002
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.002
Aromatic:NegIonizable-3.00	Number of pairs of aromatic-negIonazable atoms within 3 bonds	0.002
Aromatic:NegIonizable-6.00	Number of pairs of aromatic-negIonazable atoms	0.002

	within 6 bonds	
fr_thiophene	Number of thiophene rings	0.001
fr_HOCCN	Number of C(OH)CCN-Ctert-alkyl or C(OH)CCNcyclic	0.001
fr_tetrazole	Number of tetrazole rings	0.001
Tox_33	a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~a~3	0.001
fr_ArN	Number of N functional groups attached to aromatics	0.001

Table S11. Feature importance for predictive model of PPI inhibitors targeting the Ras/SOS1 complex.

Feature	**Description**	**Importance**
fr_aryl_methyl	Number of aryl methyl sites for hydroxylation	0.080
Acceptor:Hydrophobe-1.00	Number of pairs of acceptor-hydrophobe atoms within 1 bond	0.075
Acceptor:Hydrophobe-6.00	Number of pairs of acceptor-hydrophobe atoms within 6 bonds	0.051
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.047
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.046
Aromatic_Count	Number of atoms in aromatic rings	0.042
SMR_VSA7	MOE-type descriptors, molar refractivity and surface area contributions	0.041
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophobe atoms within 4 bonds	0.037
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.037
Donor_Count	Number of hydrogen donors	0.035
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.034
Aromatic:Aromatic-5.00	Number of pairs of aromatic-aromatic atoms within 5 bonds	0.031
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.030
fr_piperzine	Number of piperzine rings	0.030
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.029
Chi4n	Molecular connectivity index	0.029
Aromatic:Hydrophobe-5.00	Number of pairs of aromatic-hydrophobe atoms within 5 bonds	0.027
Aromatic:Aromatic-4.00	Number of pairs of aromatic-aromatic atoms within 4 bonds	0.026
fr_sulfide	Number of thioether	0.018
NegIonizable_Count	Number of negIonizable atoms	0.016
Donor:PosIonizable-5.00	Number of pairs of donor-posIonazable atoms within 5 bonds	0.016

fr_quatN	Number of quarternary nitrogens	0.014
Aromatic:Donor-1.00	Number of pairs of aromatic-donor atoms within 1 bond	0.014
fr_nitrile	Number of nitriles	0.013
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.010
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.010
Acceptor:PosIonizable-3.00	Number of pairs of acceptor-negIonazable atoms within 3 bonds	0.010
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonazable atoms within 1 bond	0.009
Aromatic:PosIonizable-2.00	Number of pairs of aromatic-posIonazable atoms within 2 bonds	0.009
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.009
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.008
PosIonizable:PosIonizable-5.00	Number of pairs of posIonazable-posIonazable atoms within 5 bonds	0.008
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.008
fr_SH	Number of thiol groups	0.007
fr_ester	Number of esters	0.007
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.007
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.006
PosIonizable:PosIonizable-6.00	Number of pairs of posIonazable-posIonazable atoms within 6 bonds	0.006
fr_ketone_Topliss	Number of ketones excluding diaryl, a,b-unsat. dienones, heteroatom on Calpha	0.006
NegIonizable:NegIonizable-3.00	Number of pairs of negIonazable-negIonazable atoms within 3 bonds	0.006
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.005
Acceptor:NegIonizable-2.00	Number of pairs of acceptor-posIonazable atoms within 2 bonds	0.005
fr_hdrzine	Number of hydrazine groups	0.005
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.005
fr_thiophene	Number of thiophene rings	0.005
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.005
fr_imidazole	Number of imidazole rings	0.004
fr_C_S	Number of thiocarbonyl	0.004
fr_urea	Number of urea groups	0.004

fr_nitro_arom	Number of nitro benzene ring substituents	0.004
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonazable atoms within 2 bonds	0.003
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.003
Aromatic:NegIonizable-3.00	Number of pairs of aromatic-negIonazable atoms within 3 bonds	0.002
Aromatic:NegIonizable-1.00	Number of pairs of aromatic-negIonazable atoms within 1 bond	0.002
fr_imide	Number of imide groups	0.002
fr_nitro_arom_nonortho	Number of non-ortho nitro benzene ring substituents	0.001

Table S12. Feature importance for predictive model of PPI inhibitors targeting the Wdr5/MLL complex.

Feature	**Description**	**Importance**
Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.344
Aromatic:PosIonizable-6.00	Number of pairs of aromatic-posIonazable atoms within 6 bonds	0.270
Chi1v	Molecular connectivity index	0.225
Acceptor:PosIonizable-3.00	Number of pairs of acceptor-negIonazable atoms within 3 bonds	0.160

Table S13. Feature importance for predictive model of PPI inhibitors targeting the Xiap/Smac complex.

Feature		**Importance**
PosIonizable_Count	Number of posIonazable atoms	0.119
fr_amide	Number of amides	0.087
Acceptor:Donor-6.00	Number of pairs of acceptor-donor atoms within 6 bonds	0.052
SlogP_VSA11	MOE-type descriptors, LogP and surface area contributions	0.035
Aromatic:Donor-5.00	Number of pairs of aromatic-donor atoms within 5 bonds	0.031
SMR_VSA9	MOE-type descriptors, molar refractivity and surface area contributions	0.027
VSA_EState8	MOE-type descriptors surface area contributions and EState indices	0.026
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.024
Aromatic_Count	Number of atoms in aromatic rings	0.024
Donor:Hydrophobe-2.00	Number of pairs of donor-hydrophobe atoms within 2 bonds	0.020
Acceptor:Acceptor-3.00	Number of pairs of acceptor-acceptor atoms within 3 bonds	0.019
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.017

PosIonizable:PosIonizable-6.00	Number of pairs of posIonazable-posIonazable atoms within 6 bonds	0.016
NOCount	Number of Nitrogens and Oxygens	0.016
fr_quatN	Number of quarternary nitrogens	0.015
PEOE_VSA8	MOE-type descriptors, partial charges and surface area contributions	0.015
NegIonizable:NegIonizable-4.00	Number of pairs of negIonazable-negIonazable atoms within 4 bonds	0.015
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.015
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.015
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.014
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.014
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.014
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-posIonazable atoms within 4 bonds	0.014
Acceptor:Hydrophobe-5.00	Number of pairs of acceptor-hydrophobe atoms within 5 bonds	0.013
Aromatic:PosIonizable-3.00	Number of pairs of aromatic-posIonazable atoms within 3 bonds	0.013
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.013
Aromatic:PosIonizable-2.00	Number of pairs of aromatic-posIonazable atoms within 2 bonds	0.013
fr_Ar_N	Number of aromatic nitrogens	0.012
NegIonizable:NegIonizable-3.00	Number of pairs of negIonazable-negIonazable atoms within 3 bonds	0.012
PEOE_VSA3	MOE-type descriptors, partial charges and surface area contributions	0.012
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.012
fr_Ar_OH	Number of aromatic hydroxyl groups	0.011
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.011
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.011
Donor:Hydrophobe-3.00	Number of pairs of donor-hydrophone atoms within 3 bonds	0.011
Acceptor:Hydrophobe-1.00	Number of pairs of acceptor-hydrophone atoms within 1 bond	0.011
Acceptor:Hydrophobe-4.00	Number of pairs of acceptor-hydrophone atoms within 4 bonds	0.011
Donor:Hydrophobe-4.00	Number of pairs of donor-hydrophone atoms within 4 bonds	0.011
Hydrophobe_Count	Number of hydrophobe atoms	0.011

Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophone atoms within 4 bonds	0.010
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophone atoms within 1 bond	0.009
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.009
Acceptor:Hydrophobe-3.00	Number of pairs of acceptor-hydrophone atoms within 3 bonds	0.009
Acceptor:Aromatic-2.00	Number of pairs of acceptor-aromatic atoms within 2 bonds	0.009
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophone atoms within 2 bonds	0.008
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.008
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.008
Aromatic:Hydrophobe-2.00	Number of pairs of aromatic-hydrophobe atoms within 2 bonds	0.008
Acceptor:PosIonizable-3.00	Number of pairs of aromatic-posIonazable atoms within 3 bonds	0.008
Hydrophobe:Hydrophobe-3.00	Number of pairs of hydrophobe-hydrophone atoms within 3 bonds	0.008
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophone atoms within 1 bond	0.007
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.007
Acceptor:NegIonizable-2.00	Number of pairs of acceptor-negIonazable atoms within 2 bonds	0.007
fr_Ar_COO	Number of Aromatic carboxylic acide	0.006
fr_oxazole	Number of oxazole rings	0.006
Acceptor:PosIonizable-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.006
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.004
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.004
Donor:NegIonizable-4.00	Number of pairs of donor-negIonazable atoms within 4 bonds	0.004
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.004
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.003
Hydrophobe:NegIonizable-2.00	Number of pairs of hydrophobe-negIonazable atoms within 2 bonds	0.003
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.003
Acceptor:PosIonizable-2.00	Number of pairs of acceptor-posIonazable atoms within 2 bonds	0.003
Donor:NegIonizable-2.00	Number of pairs of donor-posIonazable atoms	0.002

	within 2 bonds	
fr_priamide	Number of primary amides	0.002
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.002
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonazable atoms within 2 bonds	0.001
Aromatic:NegIonizable-3.00	Number of pairs of aromatic-negIonazable atoms within 3 bonds	0.001
fr_ketone_Topliss	Number of ketones excluding diaryl, a,b-unsat. dienones, heteroatom on Calpha	0.001
fr_alkyl_carbamate	Number of alkyl carbamates (subject to hydrolysis)	0.001
Aromatic:NegIonizable-1.00	Number of pairs of aromatic-negIonazable atoms within 1 bond	0.001

Table S14. Feature importance for predictive model of PPI inhibitors targeting the Annexin A2/S100-A10 complex.

Feature	Description	Importance
Aromatic:Aromatic-4.00	Number of pairs of aromatic-aromatic atoms within 4 bonds	0.569
Kappa1	Molecular shape index	0.355
fr_piperdine	Number of piperdine rings	0.076

Table S15. Feature importance for predictive model of PPI inhibitors targeting the Brd2/Ack complex.

Feature	Description	Importance
Aromatic:Aromatic-4.00	Number of pairs of aromatic-aromatic atoms within 4 bonds	0.619
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.121
Donor:Hydrophobe-2.00	Number of pairs of donor-hydrophobe atoms within 2 bonds	0.117
fr_morpholine	Number of morpholine rings	0.031
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.015
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.014
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.014
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.012
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.011
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.008
fr_alkyl_carbamate	Number of alkyl carbamates (subject to hydrolysis)	0.008
Donor:PosIonizable-1.00	Number of pairs of aromatic-aromatic atoms	0.007

	within 1 bond	
Hydrophobe:NegIonizable-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.007
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.005
Acceptor:NegIonizable-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.005
Hydrophobe:PosIonizable-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.004
Donor:NegIonizable-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.002
Aromatic:NegIonizable-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.001

Table S16. Feature importance for predictive model of PPI inhibitors targeting the CD80/CD28 complex.

Feature	**Description**	**Importance**
Tox_30	a1~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3	0.200
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.082
fr_alkyl_halide	Number of alkyl halides	0.081
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.072
Aromatic:Donor-3.00	Number of pairs of aromatic-donor atoms within 3 bonds	0.068
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.064
Aromatic:Donor-2.00	Number of pairs of aromatic-donor atoms within 2 bonds	0.056
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.050
Aromatic:Donor-1.00	Number of pairs of aromatic-donor atoms within 1 bond	0.042
Acceptor:Aromatic-3.00	Number of pairs of acceptor-aromatic atoms within 3 bonds	0.035
Acceptor:Aromatic-2.00	Number of pairs of acceptor-aromatic atoms within 2 bonds	0.024
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.023
Acceptor:Aromatic-1.00	Number of pairs of acceptor-aromatic atoms within 1 bond	0.017
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.017
Aromatic:Hydrophobe-3.00	Number of pairs of aromatic-hydrophobe atoms within 3 bonds	0.015
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.013
Donor:Hydrophobe-2.00	Number of pairs of donor-hydrophobe atoms within 2 bonds	0.011

Aromatic:Hydrophobe-2.00	Number of pairs of aromatic-hydrophobe atoms within 2 bonds	0.011
Acceptor:Donor-3.00	Number of pairs of acceptor-donor atoms within 3 bonds	0.011
Acceptor:Acceptor-2.00	Number of pairs of acceptor-acceptor atoms within 2 bonds	0.008
Hydrophobe:Hydrophobe-2.00	Number of pairs of hydrophobe-hydrophobe atoms within 2 bonds	0.007
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.006
Acceptor:Hydrophobe-3.00	Number of pairs of acceptor-hydrophobe atoms within 3 bonds	0.006
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.005
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.005
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.005
Hydrophobe:Hydrophobe-3.00	Number of pairs of hydrophobe-hydrophobe atoms within 3 bonds	0.005
Donor:Hydrophobe-3.00	Number of pairs of donor-hydrophobe atoms within 3 bonds	0.005
Acceptor:Acceptor-3.00	Number of pairs of acceptor-acceptor atoms within 3 bonds	0.005
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.004
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.004
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.003
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.003
Acceptor:Hydrophobe-1.00	Number of pairs of acceptor-hydrophobe atoms within 1 bond	0.003
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.003
Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.003
Aromatic:PosIonizable-2.00	Number of pairs of aromatic-posIonazable atoms within 2 bonds	0.002
Acceptor:Donor-2.00	Number of pairs of acceptor-donor atoms within 2 bonds	0.002
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.002
Donor:PosIonizable-3.00	Number of pairs of donor-posIonazable atoms within 3 bonds	0.002
Aromatic:PosIonizable-3.00	Number of pairs of aromatic-posIonazable atoms within 3 bonds	0.002
PosIonizable:PosIonizable-4.00	Number of pairs of posIonazable-posIonazable	0.002

	atoms within 4 bonds	
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.001
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.001
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.001
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.001
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.001
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.001
Donor:NegIonizable-2.00	Number of pairs of donor-negIonazable atoms within 2 bonds	0.001
Acceptor:PosIonizable-2.00	Number of pairs of acceptor-posIonazable atoms within 2 bonds	0.001
Donor:NegIonizable-3.00	Number of pairs of donor-negIonazable atoms within 3 bonds	0.001
Acceptor:PosIonizable-3.00	Number of pairs of acceptor-posIonazable atoms within 3 bonds	0.001

Table S17. Feature importance for predictive model of PPI inhibitors targeting the IL2/IL2-R complex.

Feature	**Description**	**Importance**
PEOE_VSA11	MOE-type descriptors, partial charges and surface area contributions	0.595
Acceptor:Aromatic-4.00	Number of pairs of acceptor-aromatic atoms within 4 bonds	0.284
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.121

Table S18. Feature importance for predictive model of PPI inhibitors targeting the Keap1/Nrf2 complex.

Feature	**Description**	**Importance**
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.192
Aromatic:NegIonizable-6.00	Number of pairs of aromatic-negIonazable atoms within 6 bonds	0.159
SMR_VSA6	MOE-type descriptors, molar refractivity and surface area contributions	0.110
Kappa3	Molecular shape index	0.098
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.091
Aromatic:Hydrophobe-6.00	Number of pairs of aromatic-hydrophobe atoms within 6 bonds	0.077
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.077
Donor:Hydrophobe-2.00	Number of pairs of donor-hydrophobe atoms	0.063

	within 2 bonds	
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonazable atoms within 2 bonds	0.044
fr_nitrile	Number of nitriles	0.031
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.028
fr_tetrazole	Number of tetrazole rings	0.012
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.005
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.004
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.003
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.003
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.001
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.001

Table S19. Feature importance for predictive model of PPI inhibitors targeting the Menin/MLL complex.

Feature	Descriptor	Importance
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.556
fr_nitrile	Number of nitriles	0.399
SlogP_VSA7	MOE-type descriptors, LogP and surface area contributions	0.045

Table S20. Feature importance for predictive model of PPI inhibitors targeting the STAT3 complex.

Feature	Descriptor	Importance
TPSA	Topological polar surface area	0.588
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.321
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.090

Table S21. Feature importance for predictive model of PPI inhibitors targeting the TTR complex.

Feature	Descriptor	Importance
Chi1n	Molecular connectivity index	0.776
SlogP_VSA7	MOE-type descriptors, LogP and surface area contributions	0.141
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.051
fr_hdrzine	Number of hydrazine groups	0.032

Table S22. Performance comparison of pdCSM-PPI and PPIM-pred on a non-redundant blind test.

	pdCSM-PPI				PPIM-pred			
PPI target	**F1**	**AUC**	**Sensitivity**	**Specificity**	**F1**	**AUC**	**Sensitivity**	**Specificity**
Bcl2-Like / Bak-Bax	0.99	1.00	1.00	0.98	0.26	0.67+	0.71*	0.67*
Mdm2-Like / P53	0.97	1.00	0.97	0.96	0.45	0.84+	1.00*	0.79*

* p-value < 0.05 by McNemar's Test compared to pdCSM-PPI

+ p-value < 0.05 by z transformation test compared to pdCSM-PPI

Table S23. Performance comparison of pdCSM-PPI and SMMPPI on non-redundant blind tests.

	pdCSM-PPI			SMMPPI		
PPI target	**MCC**	**F1**	**AUC**	**MCC**	**F1**	**AUC**
Bcl2-Like / Bak-Bax	0.86	0.92	0.92	0.70	0.84	0.92
Bromodomain / Histone	0.82	0.90	0.90	0.73	0.87	0.93
LEDGF / IN	0.70	0.85	0.85	0.70	0.85	0.97
LFA / ICAM	1.00	1.00	1.00	0.92	0.96	0.99
Mdm2-Like / P53	0.88	0.93	0.93	0.79	0.89	0.93
Ras / SOS1	0.57	0.73	0.75	0.52	0.71	0.82
XIAP / Smac	0.89	0.94	0.94	0.89	0.95	0.97
WDR5 / MLL	0.92	0.96	0.94	0.74	0.87	0.87
CD4 / gp120	0.91	0.95	0.95	0.90	0.95	0.95

Table S24. Distribution of compounds with experimentally determined inhibitory activity (IC50) across 45 different PPIs, on the original dataset retrieved from TIMBAL and iPPI-DB, and after clustering using the Butina algorithm with Tanimoto similarity cutoff of 0.8. PPI targets with more than 40 inhibitors after clustering were further selected for building regression models.

PPI	# compounds	# compounds (clustering 0.8)
Integrins	1606	907
Mdm2-Like / P53	685	400
LFA / ICAM	277	150
BCL2-Like / BAX-BAK	350	127
Bromodomain / Histone	145	111
HIF-1a / p300	121	82
Cyclophilins	105	73
LEDGF / IN	74	68
XIAP / Smac	69	42
CD80 / CD28	73	35
BRD2 / Ack	47	32
Annexin A2/S100-A10	44	29
Neuropilin / VEGF	41	24
FKBP1A/FK506	29	21
STAT3	38	21
TTR	19	17

Transthyretin	16	16
BRD4 / NUT	16	15
BetaCatenin / Tcf4 and Tcf3	40	15
IL2 / IL-2R	29	15
Rac1	15	15
Tubulin	14	14
MENIN / MLL	20	12
SPIN1 / H3	15	10
E2 / E1	10	8
MLLT1 / H3	16	8
PCNA trimer	7	7
c-Myc / Max	8	7
SOD1	10	5
RUNX1 / CBFb	4	4
VEGF / VEGFR	4	4
53BP1 / H4	3	3
WD40 / H3	3	3
CRM1 / Rev	2	2
NRP / VEGF	2	2
Ras / SOS1	2	2
UPAR / UPA	3	2
WDR5/MLL	3	2
BRI1	1	1
CD4 / gp120	1	1
CD40 / CD154	1	1
CaM / CaMBD2	1	1
Rad51	1	1
TNFa / TNFa	1	1
Tak1 / Tab1	1	1

Table S25. Feature importance for the regression model of PPI inhibitors targeting the BCL2-Like / BAX-BAK complex.

Feature	**Description**	**Importance**
SMR_VSA5	MOE-type descriptors, molar refractivity and surface area contributions	0.128
SlogP_VSA7	MOE-type descriptors, LogP and surface area contributions	0.104
Chi2n	Molecular connectivity index	0.096
fr_Ar_N	Number of aromatic nitrogens	0.083
fr_bicyclic	Number of bicyclic structures	0.042
Donor_Count	Number of hydrogen donors	0.042
SMR_VSA10	MOE-type descriptors, molar refractivity and	0.037

	surface area contributions	
VSA_EState8	MOE-type descriptors, surface area contributions and EState indices	0.036
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.034
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.032
Acceptor:Hydrophobe-3.00	Number of pairs of acceptor-hydrophobe atoms within 3 bonds	0.032
PEOE_VSA9	MOE-type descriptors, partial charges and surface area contributions	0.028
fr_NH2	Number of Primary amines	0.028
Aromatic_Count	Number of atoms in aromatic rings	0.025
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.024
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.023
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophobe atoms within 4 bonds	0.021
Donor:Hydrophobe-3.00	Number of pairs of donor-hydrophobe atoms within 3 bonds	0.021
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.020
fr_thiazole	Number of thiazole rings	0.016
SMR_VSA9	MOE-type descriptors, molar refractivity and surface area contributions	0.013
SlogP_VSA12	MOE-type descriptors, LogP and surface area contributions	0.010
fr_unbrch_alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)	0.008
fr_alkyl_halide	Number of alkyl halides	0.007
Acceptor:PosIonizable-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.007
Donor:PosIonizable-3.00	Number of pairs of donor-posIonazable atoms within 3 bonds	0.006
PosIonizable_Count	Number of posIonazable atoms	0.006
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.006
PosIonizable:PosIonizable-4.00	Number of pairs of posIonazable-posIonazable atoms within 4 bonds	0.005
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-posIonazable atoms within 4 bonds	0.005
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.005
Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.004
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.004

fr_sulfide	Number of thioether	0.004
fr_Ar_OH	Number of aromatic hydroxyl groups	0.003
Acceptor:NegIonizable-4.00	Number of pairs of acceptor-negIonazable atoms within 4 bonds	0.003
fr_pyridine	Number of pyridine rings	0.003
Acceptor:NegIonizable-2.00	Number of pairs of acceptor-negIonazable atoms within 2 bonds	0.003
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.003
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.003
NegIonizable:NegIonizable-5.00	Number of pairs of negIonazable-negIonazable atoms within 5 bonds	0.002
NegIonizable:NegIonizable-4.00	Number of pairs of negIonazable-negIonazable atoms within 4 bonds	0.002
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonazable atoms within 1 bond	0.002
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.002
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.002
Hydrophobe:NegIonizable-4.00	Number of pairs of hydrophobe-negIonazable atoms within 4 bonds	0.002
Donor:NegIonizable-3.00	Number of pairs of donor-negIonazable atoms within 3 bonds	0.002
Donor:NegIonizable-5.00	Number of pairs of donor-negIonazable atoms within 5 bonds	0.002
fr_phenol	Number of phenols	0.002
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.001
fr_ketone	Number of ketones	0.001
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.001
Aromatic:NegIonizable-4.00	Number of pairs of aromatic-negIonazable atoms within 4 bonds	0.001
Hydrophobe:NegIonizable-2.00	Number of pairs of hydrophobe-negIonazable atoms within 2 bonds	0.001

Table S26. Feature importance for the regression model of PPI inhibitors targeting the Bromodomain / Histone complex.

Feature	**Description**	**Importance**
Chi1v	Molecular connectivity index	0.219
BalabanJ	Balaban's connectivity topological index	0.083
Kappa2	Molecular shape index	0.075
Tox_2	a[NH2]	0.054
PEOE_VSA7	MOE-type descriptors, partial charges and surface area contributions	0.039

PEOE_VSA9	MOE-type descriptors, partial charges and surface area contributions	0.039
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.026
SMR_VSA4	MOE-type descriptors, molar refractivity and surface area contributions	0.025
Acceptor:Acceptor-5.00	Number of pairs of acceptor-acceptor atoms within 5 bonds	0.025
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophobe atoms within 4 bonds	0.024
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.022
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.020
PEOE_VSA2	MOE-type descriptors, partial charges and surface area contributions	0.018
Aromatic:Hydrophobe-5.00	Number of pairs of aromatic-hydrophobe atoms within 5 bonds	0.018
fr_NH0	Number of Tertiary amines	0.017
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.017
VSA_EState10	MOE-type descriptors surface area contributions and EState indices	0.017
fr_benzene	Number of benzene rings	0.015
fr_Imine	Number of Imines	0.014
SMR_VSA6	MOE-type descriptors, molar refractivity and surface area contributions	0.014
Aromatic_Count	Number of atoms in aromatic rings	0.013
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.011
SlogP_VSA1	MOE-type descriptors, LogP and surface area contributions	0.011
SMR_VSA5	MOE-type descriptors, molar refractivity and surface area contributions	0.011
fr_aniline	Number of anilines	0.011
Aromatic:Aromatic-5.00	Number of pairs of aromatic-aromatic atoms within 5 bonds	0.010
Acceptor:Aromatic-2.00	Number of pairs of acceptor-aromatic atoms within 2 bonds	0.010
VSA_EState8	MOE-type descriptors, surface area contributions and EState indices	0.009
Acceptor:Aromatic-5.00	Number of pairs of acceptor-aromatic atoms within 5 bonds	0.009
Acceptor:Aromatic-4.00	Number of pairs of acceptor-aromatic atoms within 4 bonds	0.009
PEOE_VSA3	MOE-type descriptors, partial charges and surface area contributions	0.009
fr_aryl_methyl	Number of aryl methyl sites for hydroxylation	0.009
Donor:Hydrophobe-4.00	Number of pairs of donor-hydrophobe atoms	0.008

	within 4 bonds	
Aromatic:Donor-3.00	Number of pairs of aromatic-donor atoms within 3 bonds	0.007
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.007
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.006
fr_thiophene	Number of thiophene rings	0.006
Acceptor:PosIonizable-6.00	Number of pairs of acceptor-posIonazable atoms within 6 bonds	0.006
PEOE_VSA11	MOE-type descriptors, partial charges and surface area contributions	0.006
PEOE_VSA5	MOE-type descriptors, partial charges and surface area contributions	0.006
Aromatic:PosIonizable-6.00	Number of pairs of aromatic-posIonazable atoms within 6 bonds	0.005
Aromatic:Donor-1.00	Number of pairs of aromatic-donor atoms within 1 bond	0.005
PEOE_VSA4	MOE-type descriptors, partial charges and surface area contributions	0.004
Aromatic:PosIonizable-5.00	Number of pairs of aromatic-posIonazable atoms within 5 bonds	0.004
fr_Ar_OH	Number of aromatic hydroxyl groups	0.003
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.003
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.002
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.002
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.002
PosIonizable_Count	Number of posIonazable atoms	0.002
fr_halogen	Number of halogens	0.002
Donor:Donor-4.00	Number of pairs of donor-donor atoms within 4 bonds	0.002
NegIonizable_Count	Number of negIonazable atoms	0.001
Acceptor:PosIonizable-2.00	Number of pairs of acceptor-posIonazable atoms within 2 bonds	0.001
Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.001
Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.001
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.001
Donor:PosIonizable-4.00	Number of pairs of donor-posIonazable atoms within 4 bonds	0.001
Aromatic:PosIonizable-4.00	Number of pairs of aromatic-posIonazable atoms within 4 bonds	0.001

PosIonizable:PosIonizable-5.00	Number of pairs of posIonazable-posIonazable atoms within 5 bonds	0.001
Acceptor:PosIonizable-3.00	Number of pairs of acceptor-posIonazable atoms within 3 bonds	0.001

Table S27. Feature importance for the regression model of PPI inhibitors targeting the HIF-1a / p300 complex.

Feature	Description	Importance
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.464
RingCount	Number of rings	0.192
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.09
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.076
fr_Al_OH	Number of aliphatic hydroxyl groups	0.044
Donor:PosIonizable-6.00	Number of pairs of donor-posIonazable atoms within 6 bonds	0.044
VSA_EState10	MOE-type descriptors, surface area contributions and EState indices	0.041
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.025
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.012
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.006
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.004
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.001
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 1 bonds	0.001

Table S28. Feature importance for the regression model of PPI inhibitors targeting the Mdm2-Like / P53 complex.

Feature	Description	Importance
PEOE_VSA13	MOE-type descriptors, partial charges and surface area contributions	0.058
Chi4n	Molecular connectivity index	0.054
VSA_EState8	MOE-type descriptors, surface area contributions and EState indices	0.051
fr_halogen	Number of halogens	0.046
VSA_EState10	MOE-type descriptors, surface area contributions and EState indices	0.035
PEOE_VSA9	MOE-type descriptors, partial charges and	0.035

	surface area contributions	
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.031
fr_bicyclic	Number of bicyclic structures	0.030
fr_NH1	Number of Secondary amines	0.029
Chi0n	Molecular connectivity index	0.026
BalabanJ	Balaban's connectivity topological index	0.023
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.023
Chi3n	Molecular connectivity index	0.023
PEOE_VSA11	MOE-type descriptors, partial charges and surface area contributions	0.021
VSA_EState9	MOE-type descriptors surface area contributions and EState indices	0.020
Kappa1	Molecular shape index	0.018
Donor_Count	Number of hydrogen donors	0.018
HeavyAtomCount	Number of heavy atoms	0.018
fr_aniline	Number of anilines	0.017
FCount	Number of fluorine atoms	0.016
MolWt	Molecular weight	0.015
PEOE_VSA8	MOE-type descriptors, partial charges and surface area contributions	0.015
SMR_VSA7	MOE-type descriptors, molar refractivity and surface area contributions	0.014
Acceptor:Hydrophobe-4.00	Number of pairs of acceptor-hydrophobe atoms within 4 bonds	0.012
Donor:Donor-4.00	Number of pairs of donor-donor atoms within 4 bonds	0.012
Donor:Hydrophobe-4.00	Number of pairs of donor-hydrophobe atoms within 4 bonds	0.012
fr_NH2	Number of Primary amines	0.011
fr_C_O_noCOO	Number of carbonyl O, excluding COOH	0.011
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.011
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.011
SMR_VSA1	MOE-type descriptors, molar refractivity and surface area contributions	0.011
fr_C_O	Number of carbonyl O	0.011
Donor:NegIonizable-1.00	Number of pairs of donor-negIonazable atoms within 1 bond	0.010
Acceptor:Acceptor-3.00	Number of pairs of acceptor-acceptor atoms within 3 bonds	0.010
Acceptor:Aromatic-1.00	Number of pairs of acceptor-aromatic atoms within 1 bond	0.010
fr_HOCCN	Number of C(OH)CCN-Ctert-alkyl or	0.009

	C(OH)CCNcyclic	
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.009
Acceptor:Hydrophobe-6.00	Number of pairs of acceptor-hydrophobe atoms within 6 bonds	0.009
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.008
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.008
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.008
PEOE_VSA14	MOE-type descriptors, partial charges and surface area contributions	0.008
fr_aryl_methyl	Number of aryl methyl sites for hydroxylation	0.008
Acceptor:Aromatic-3.00	Number of pairs of acceptor-aromatic atoms within 3 bonds	0.008
SMR_VSA9	MOE-type descriptors, molar refractivity and surface area contributions	0.007
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.006
NegIonizable:NegIonizable-2.00	Number of pairs of negIonazable-negIonazable atoms within 2 bonds	0.006
Donor:PosIonizable-5.00	Number of pairs of donor-posIonazable atoms within 5 bonds	0.006
NegIonizable:NegIonizable-3.00	Number of pairs of negIonazable-negIonazable atoms within 3 bonds	0.006
fr_Ndealkylation1	Number of XCCNR groups	0.006
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.006
Aromatic:Aromatic-4.00	Number of pairs of aromatic-aromatic atoms within 4 bonds	0.005
Aromatic:Aromatic-5.00	Number of pairs of aromatic-aromatic atoms within 5 bonds	0.005
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.005
Aromatic:Aromatic-6.00	Number of pairs of aromatic-aromatic atoms within 6 bonds	0.005
Donor:NegIonizable-2.00	Number of pairs of donor-negIonazable atoms within s bonds	0.005
Acceptor:PosIonizable-5.00	Number of pairs of acceptor-posIonazable atoms within 5 bonds	0.005
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.005
Aromatic_Count	Number of atoms in aromatic rings	0.005
fr_piperzine	Number of piperzine rings	0.004
NegIonizable:NegIonizable-6.00	Number of pairs of negIonazable-negIonazable atoms within 6 bonds	0.004
Donor:PosIonizable-3.00	Number of pairs of donor-posIonazable atoms within 3 bonds	0.004

fr_pyridine	Number of pyridine rings	0.004
NegIonizable:NegIonizable-1.00	Number of pairs of negIonazable-negIonazable atoms within 1 bond	0.004
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bonds	0.004
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.004
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.004
fr_ArN	Number of N functional groups attached to aromatics	0.004
Aromatic:PosIonizable-3.00	Number of pairs of aromatic-posIonazable atoms within 3 bonds	0.004
Hydrophobe:PosIonizable-3.00	Number of pairs of hydrophobe-posIonazable atoms within 3 bonds	0.004
fr_nitro	Number of nitro groups	0.003
fr_morpholine	Number of morpholine rings	0.003
Hydrophobe:NegIonizable-1.00	Number of pairs of hydrophobe-negIonazable atoms within 1 bond	0.003
Tox_1	O=N(~O)a	0.003
fr_nitro_arom	Number of nitro benzene ring substituents	0.003
fr_imide	Number of imide groups	0.002
Acceptor:NegIonizable-1.00	Number of pairs of acceptor-negIonazable atoms within 1 bond	0.002
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-posIonazable atoms within 4 bonds	0.002
Aromatic:NegIonizable-6.00	Number of pairs of aromatic-negIonazable atoms within 63 bonds	0.002
fr_nitrile	Number of nitriles	0.002
fr_thiophene	Number of thiophene rings	0.001
fr_N_O	Number of hydroxylamine groups	0.001
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonazable atoms within 2 bonds	0.001
Aromatic:NegIonizable-4.00	Number of pairs of aromatic-negIonazable atoms within 4 bonds	0.001
Aromatic:NegIonizable-3.00	Number of pairs of aromatic-negIonazable atoms within 3 bonds	0.001

Table S29. Feature importance for the regression model of PPI inhibitors targeting the Integrins complex.

Feature	**Description**	**Importance**
fr_Al_COO	Number of aliphatic carboxylic acids	0.067
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.059
fr_pyridine	Number of pyridine rings	0.055

Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe-hydrophobe atoms within 4 bonds	0.049
MolWt	Molecular weight	0.044
SlogP_VSA2	MOE-type descriptors, LogP and surface area contributions	0.042
Acceptor:Aromatic-5.00	Number of pairs of acceptor-aromatic atoms within 5 bonds	0.040
HallKierAlpha	Hall-Kier alpha value	0.038
Aromatic_Count	Number of atoms in aromatic rings	0.034
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.033
Acceptor:Acceptor-6.00	Number of pairs of acceptor-acceptor atoms within 6 bonds	0.033
fr_NH1	Number of Secondary amines	0.032
fr_piperdine	Number of piperdine rings	0.030
SlogP_VSA4	MOE-type descriptors, LogP and surface area contributions	0.027
PEOE_VSA2	MOE-type descriptors, partial charges and surface area contributions	0.026
PEOE_VSA12	MOE-type descriptors, partial charges and surface area contributions	0.026
PEOE_VSA14	MOE-type descriptors, partial charges and surface area contributions	0.023
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.023
Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.021
Hydrophobe:NegIonizable-2.00	Number of pairs of hydrophobe-negIonazable atoms within 2 bonds	0.018
Donor:Donor-4.00	Number of pairs of donor-donor atoms within 4 bonds	0.018
PEOE_VSA13	MOE-type descriptors, partial charges and surface area contributions	0.018
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.018
Hydrophobe:NegIonizable-3.00	Number of pairs of hydrophobe-negIonazable atoms within 3 bonds	0.017
Hydrophobe:NegIonizable-4.00	Number of pairs of hydrophobe-negIonazable atoms within 4 bonds	0.017
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.014
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.013
fr_amidine	Number of amidine groups	0.012
Donor:NegIonizable-2.00	Number of pairs of donor-negIonazable atoms within 2 bonds	0.012
fr_sulfonamd	Number of sulfonamides	0.011
Hydrophobe:PosIonizable-5.00	Number of pairs of hydrophober-posIonazable	0.011

	atoms within 5 bonds	
Acceptor:PosIonizable-6.00	Number of pairs of acceptor-posIonazable atoms within 6 bonds	0.010
fr_Imine	Number of Imines	0.009
PosIonizable:PosIonizable-6.00	Number of pairs of posIonazable-posIonazable atoms within 6 bonds	0.008
fr_alkyl_halide	Number of alkyl halides	0.007
fr_Ndealkylation1	Number of XCCNR groups	0.007
Aromatic:PosIonizable-6.00	Number of pairs of aromatic-posIonazable atoms within 6 bonds	0.007
Donor:PosIonizable-6.00	Number of pairs of donor-posIonazable atoms within 6 bonds	0.006
Donor:PosIonizable-5.00	Number of pairs of donor-posIonazable atoms within 5 bonds	0.006
PosIonizable:PosIonizable-5.00	Number of pairs of posIonazable-posIonazable atoms within 5 bonds	0.006
NegIonizable:PosIonizable-6.00	Number of pairs of negIonazable-posIonazable atoms within 6 bonds	0.005
FCount	Number of fluorine atoms	0.005
fr_ArN	Number of N functional groups attached to aromatics	0.005
Donor:PosIonizable-4.00	Number of pairs of donor-posIonazable atoms within 4 bonds	0.004
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.004
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.004
Acceptor:PosIonizable-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.004
Acceptor:PosIonizable-1.00	Number of pairs of acceptor-posIonazable atoms within 1 bond	0.003
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.003
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.003
Aromatic:PosIonizable-2.00	Number of pairs of aromatic-posIonazable atoms within 2 bonds	0.002
NegIonizable:PosIonizable-2.00	Number of pairs of negIonazable-posIonazable atoms within 2 bonds	0.002
fr_nitrile	Number of nitriles	0.002
fr_imide	Number of imide groups	0.002
fr_barbitur	Number of barbiturate groups	0.001
Hydrophobe:PosIonizable-1.00	Number of pairs of hydrophobe-posIonazable atoms within 1 bond	0.001
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.001
fr_term_acetylene	Number of terminal acetylenes	0.001

Table S30. Feature importance for the regression model of PPI inhibitors targeting the LFA / ICAM complex.

Feature	Description	Importance
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.149
VSA_EState8	MOE-type descriptors surface area contributions and EState indices	0.145
Kappa2	Molecular shape index	0.093
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.091
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.091
fr_bicyclic	Number of bicyclic structures	0.082
SMR_VSA5	MOE-type descriptors, molar refractivity and surface area contributions	0.078
Acceptor:Donor-4.00	Number of pairs of acceptor-donor atoms within 4 bonds	0.072
fr_para_hydroxylation	Number of para-hydroxylation sites	0.068
fr_urea	Number of urea groups	0.043
fr_imide	Number of imide groups	0.040
fr_amide	Number of amides	0.035
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.012

Table S31. Feature importance for the regression model of PPI inhibitors targeting the Cyclophilins complex.

Feature	Description	Importance
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.176
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posgIonazable atoms within 2 bonds	0.135
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.113
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.078
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.075
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.071
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.070
Acceptor:Acceptor-4.00	Number of pairs of acceptor-acceptor atoms within 4 bonds	0.055
Acceptor:Acceptor-2.00	Number of pairs of acceptor-acceptor atoms within 2 bonds	0.047

Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.040
SMR_VSA6	MOE-type descriptors, molar refractivity and surface area contributions	0.019
Chi1n	Molecular connectivity index	0.019
fr_amide	Number of amides	0.015
Acceptor:Donor-1.00	Number of pairs of acceptor-donor atoms within 1 bond	0.012
Donor:Hydrophobe-1.00	Number of pairs of donor-hydrophobe atoms within 1 bond	0.009
Aromatic:Donor-1.00	Number of pairs of aromatic-donor atoms within 1 bond	0.009
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.008
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.008
Acceptor:Acceptor-1.00	Number of pairs of acceptor-acceptor atoms within 1 bond	0.006
Hydrophobe:Hydrophobe-2.00	Number of pairs of hydrophobe-hydrophobe atoms within 2 bonds	0.005
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.004
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.004
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.004
Aromatic:Hydrophobe-2.00	Number of pairs of aromatic-hydrophobe atoms within 2 bonds	0.004
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.004
fr_benzene	Number of benzene rings	0.003
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.003
fr_pyridine	Number of pyridine rings	0.001
fr_Ar_N	Number of aromatic nitrogens	0.001

Table S32. Feature importance for the regression model of PPI inhibitors targeting the LEDGF / IN complex.

Feature	**Description**	**Importance**
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.132
Acceptor:Acceptor-5.00	Number of pairs of acceptor-acceptor atoms within 5 bonds	0.065
PEOE_VSA10	MOE-type descriptors, partial charges and surface area contributions	0.059
SlogP_VSA2	MOE-type descriptors, LogP and surface area contributions	0.041

PEOE_VSA6	MOE-type descriptors, partial charges and surface area contributions	0.039
Kappa1	Molecular shape index	0.034
PEOE_VSA7	MOE-type descriptors, partial charges and surface area contributions	0.029
SlogP_VSA4	MOE-type descriptors, LogP and surface area contributions	0.027
Chi1v	Molecular connectivity index	0.027
PEOE_VSA5	MOE-type descriptors, partial charges and surface area contributions	0.027
Chi0v	Molecular connectivity index	0.026
Chi2n	Molecular connectivity index	0.025
PEOE_VSA2	MOE-type descriptors, partial charges and surface area contributions	0.024
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.022
fr_nitro_arom	Number of nitro benzene ring substituents	0.021
Donor:NegIonizable-4.00	Number of pairs of donor-negIonazable atoms within 4 bonds	0.021
fr_C_O_noCOO	Number of carbonyl O, excluding COOH	0.021
SMR_VSA3	MOE-type descriptors, molar refractivity and surface area contributions	0.020
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.020
Chi0	Molecular connectivity index	0.019
Tox_1	O=N(~O)a	0.019
Donor:Donor-5.00	Number of pairs of donor-donor atoms within 5 bonds	0.018
fr_nitro	Number of nitro groups	0.018
fr_hdrzone	Number of hydrazone groups	0.018
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.017
Donor:Hydrophobe-5.00	Number of pairs of donor-hydrophobe atoms within 5 bonds	0.017
Donor:NegIonizable-3.00	Number of pairs of donor-negIonazable atoms within 3 bonds	0.016
HallKierAlpha	Hall-Kier alpha value	0.015
Acceptor:Donor-2.00	Number of pairs of acceptor-donor atoms within 2 bonds	0.015
Aromatic:Donor-3.00	Number of pairs of aromatic-donor atoms within 3 bonds	0.014
Donor:Donor-1.00	Number of pairs of donor-donor atoms within 1 bond	0.012
PEOE_VSA12	MOE-type descriptors, partial charges and surface area contributions	0.012
Aromatic:Hydrophobe-4.00	Number of pairs of aromatic-hydrophobe atoms within 4 bonds	0.010

Acceptor:Donor-5.00	Number of pairs of acceptor-donor atoms within 5 bonds	0.010
fr_pyridine	Number of pyridine rings	0.010
Acceptor:Aromatic-4.00	Number of pairs of acceptor-aromatic atoms within 4 bonds	0.010
Hydrophobe:Hydrophobe-3.00	Number of pairs of hydrophobe-hydrophobe atoms within 3 bonds	0.009
fr_ArN	Number of N functional groups attached to aromatics	0.007
Acceptor:Hydrophobe-5.00	Number of pairs of acceptor-hydrophobe atoms within 5 bonds	0.007
SlogP_VSA7	MOE-type descriptors, LogP and surface area contributions	0.007
Acceptor:Hydrophobe-4.00	Number of pairs of acceptor-hydrophobe atoms within 4 bonds	0.006
fr_amide	Number of amides	0.005
Tox_2	a[NH2]	0.003
PosIonizable:PosIonizable-1.00	Number of pairs of posIonazable-posIonazable atoms within 1 bond	0.003
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.003
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.002
Tox_12	[OH,NH2][N,O]	0.002
fr_furan	Number of furan rings	0.002
Donor:PosIonizable-1.00	Number of pairs of donor-posIonazable atoms within 1 bond	0.002
fr_azo	Number of azo groups	0.002
fr_ketone	Number of ketones	0.002
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.002
Aromatic:PosIonizable-1.00	Number of pairs of aromatic-posIonazable atoms within 1 bond	0.002
fr_thiazole	Number of thiazole rings	0.002
fr_unbrch_alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)	0.001
fr_C_S	Number of thiocarbonyl	0.001
fr_amidine	Number of amidine groups	0.001
fr_Al_OH_noTert	Number of aliphatic hydroxyl groups excluding tert-OH	0.001

Table S33. Feature importance for the regression model of PPI inhibitors targeting the XIAP / Smac complex.

Feature	**Description**	**Importance**
fr_Nhpyrrole	Number of H-pyrrole nitrogens	0.120
Acceptor:Hydrophobe-5.00	Number of pairs of acceptor-hydrophobe	0.056

	atoms within 5 bonds	
SMR_VSA9	MOE-type descriptors, molar refractivity and surface area contributions	0.055
Acceptor:Donor-6.00	Number of pairs of acceptor-donor atoms within 6 bonds	0.054
Donor:Donor-3.00	Number of pairs of donor-donor atoms within 3 bonds	0.052
Donor:Donor-2.00	Number of pairs of donor-donor atoms within 2 bonds	0.050
NOCount	Number of Nitrogens and Oxygens	0.044
Hydrophobe:PosIonizable-2.00	Number of pairs of hydrophobe-posIonazable atoms within 2 bonds	0.042
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.040
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-posIonazable atoms within 4 bonds	0.036
Aromatic:Aromatic-2.00	Number of pairs of aromatic-aromatic atoms within 2 bonds	0.031
Donor:Hydrophobe-3.00	Number of pairs of donor-hydrophobe atoms within 3 bonds	0.030
Donor:Hydrophobe-4.00	Number of pairs of donor-hydrophobe atoms within 4 bonds	0.027
Aromatic:Aromatic-1.00	Number of pairs of aromatic-aromatic atoms within 1 bond	0.026
SMR_VSA10	MOE-type descriptors, molar refractivity and surface area contributions	0.025
Aromatic:Donor-5.00	Number of pairs of aromatic-donor atoms within 5 bonds	0.021
Hydrophobe_Count	Number of hydrophobe atoms	0.021
Acceptor:Aromatic-2.00	Number of pairs of acceptor-aromatic atoms within 2 bonds	0.021
fr_priamide	Number of primary amides	0.020
Acceptor:Hydrophobe-4.00	Number of pairs of acceptor-hydrophobe atoms within 4 bonds	0.020
fr_amide	Number of amides	0.018
Aromatic:Hydrophobe-2.00	Number of pairs of aromatic-hydrophobe atoms within 2 bonds	0.018
Hydrophobe:Hydrophobe-1.00	Number of pairs of hydrophobe-hydrophobe atoms within 1 bond	0.018
Hydrophobe:Hydrophobe-3.00	Number of pairs of hydrophobe-hydrophobe atoms within 3 bonds	0.018
Aromatic_Count	Number of atoms in aromatic rings	0.017
Acceptor:Hydrophobe-3.00	Number of pairs of acceptor-hydrophobe atoms within 3 bonds	0.016
Hydrophobe:Hydrophobe-4.00	Number of pairs of hydrophobe atoms within 4 bonds	0.015
Aromatic:Hydrophobe-1.00	Number of pairs of aromatic-hydrophobe atoms within 1 bond	0.011

Acceptor:PosIonizable-2.00	Number of pairs of acceptor-posIonazable atoms within 2 bonds	0.011
PosIonizable_Count	Number of posIonazable atoms	0.010
PosIonizable:PosIonizable-3.00	Number of pairs of posIonazable-posIonazable atoms within 3 bonds	0.008
Acceptor:PosIonizable-3.00	Number of pairs of acceptor-posIonazable atoms within 3 bonds	0.006
PEOE_VSA3	MOE-type descriptors, partial charges and surface area contributions	0.006
PEOE_VSA8	MOE-type descriptors, partial charges and surface area contributions	0.005
VSA_EState8	MOE-type descriptors, surface area contributions and EState indices	0.005
SlogP_VSA11	MOE-type descriptors, LogP and surface area contributions	0.005
fr_Ar_N	Number of aromatic nitrogens	0.004
Acceptor:Hydrophobe-2.00	Number of pairs of acceptor-hydrophobe atoms within 2 bonds	0.004
Acceptor:Hydrophobe-1.00	Number of pairs of acceptor-hydrophobe atoms within 1 bond	0.004
PosIonizable:PosIonizable-6.00	Number of pairs of posIonazable-posIonazable atoms within 6 bonds	0.004
Donor:Hydrophobe-2.00	Number of pairs of donor-hydrophobe atoms within 2 bonds	0.002
Acceptor:Acceptor-3.00	Number of pairs of acceptor-acceptor atoms within 3 bonds	0.002
Acceptor:PosIonizable-4.00	Number of pairs of acceptor-posIonazable atoms within 4 bonds	0.001

Table S34. Feature importance for general predictive model of PPI inhibitors activity (IC50).

Feature	**Description**	**Importance**
BalabanJ	Balaban's connectivity topological index	0.071
NumHeteroatoms	Number of heteroatoms	0.057
fr_C_O	Number of carbonyl O	0.046
Chi4v	Molecular connectivity index	0.044
SlogP_VSA5	MOE-type descriptors, LogP and surface area contributions	0.04
fr_amide	Number of amides	0.035
PEOE_VSA3	MOE-type descriptors, partial charges and surface area contributions	0.033
PEOE_VSA10	MOE-type descriptors, partial charges and surface area contributions	0.032
Acceptor:Aromatic-6.00	Number of pairs of aromatic-aromatic atoms within 6 bonds	0.031
Aromatic:NegIonizable-2.00	Number of pairs of aromatic-negIonazable atoms within 2 bonds	0.031
Hydrophobe:Hydrophobe-6.00	Number of pairs of hydrophobe-hydrophobe	0.029

	atoms within 6 bonds	
Donor:Donor-6.00	Number of pairs of donor-donor atoms within 6 bonds	0.029
SMR_VSA7	MOE-type descriptors, molar refractivity and surface area contributions	0.029
SlogP_VSA10	MOE-type descriptors, LogP and surface area contributions	0.028
Donor_Count	Number of hydrogen donors	0.028
PEOE_VSA8	MOE-type descriptors, partial charges and surface area contributions	0.028
fr_NH0	Number of Tertiary amines	0.027
Hydrophobe_Count	Number of hydrophobe atoms	0.027
Hydrophobe:Hydrophobe-5.00	Number of pairs of hydrophobe-hydrophobe atoms within 5 bonds	0.027
fr_bicyclic	Number of bicyclic structures	0.023
SlogP_VSA12	MOE-type descriptors, LogP and surface area contributions	0.022
Aromatic:Aromatic-3.00	Number of pairs of aromatic-aromatic atoms within 3 bonds	0.02
Aromatic:Aromatic-6.00	Number of pairs of aromatic-aromatic atoms within 6 bonds	0.02
PosIonizable:PosIonizable-2.00	Number of pairs of posIonazable-posIonazable atoms within 2 bonds	0.019
fr_methoxy	Number of methoxy groups -OCH3	0.019
Aromatic:Aromatic-5.00	Number of pairs of aromatic-aromatic atoms within 5 bonds	0.019
SlogP_VSA8	MOE-type descriptors, LogP and surface area contributions	0.018
fr_amidine	Number of amidine groups	0.018
fr_unbrch_alkane	Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)	0.017
Donor:PosIonizable-6.00	Number of pairs of donor-posIonazable atoms within 6 bonds	0.014
Hydrophobe:PosIonizable-4.00	Number of pairs of hydrophobe-posIonazable atoms within 4 bonds	0.013
fr_phenol	Number of phenols	0.013
fr_ketone	Number of ketones	0.013
Hydrophobe:PosIonizable-6.00	Number of pairs of hydrophobe-posIonazable atoms within 6 bonds	0.012
fr_ester	Number of esters	0.011
fr_phos_ester	Number of phosphoric ester groups	0.009
fr_guanido	Number of guanidine groups	0.009
fr_sulfide	Number of thioether	0.007
Aromatic:PosIonizable-6.00	Number of pairs of aromatic-posIonazable atoms within 6 bonds	0.007
Aromatic:PosIonizable-4.00	Number of pairs of aromatic-posIonazable atoms within 4 bonds	0.006

fr_phos_acid	Number of phosphoric acid groups	0.006
Donor:PosIonizable-3.00	Number of pairs of donor-posIonazable atoms within 3 bonds	0.006
Donor:PosIonizable-2.00	Number of pairs of donor-posIonazable atoms within 2 bonds	0.006
fr_term_acetylene	Number of terminal acetylenes	0.002

Table S35. Distribution of PPIs inhibitors retrieved from 3 databases: TIMBAL, iPPI-DB, 2P2I-DB. Number of compounds per PPI target is shown before and after clustering with Tanimoto similarity cutoff of 0.8.

PPI	**# compounds**	**# compounds (clustering 0.8)**
Integrins	1606	907
Bromodomain / Histone	680	501
MDM2-Like / P53	768	440
LFA / ICAM	277	150
BCL2-Like / BAX-BAK	386	149
LEDGF / IN	158	122
HIF-1a	121	82
Cyclophilins	105	73
XIAP / Smac	99	68
Ras / SOS1	64	55
WDR5/MLL	39	37
CD80 / CD28	73	35
BRD2 / Ack	47	32
MENIN / MLL	49	32
Annexin A2/S100-A10	44	29
Keap1 / Nrf2	31	26
Neuropilin / VEGF	41	24
FKBP1A/FK506	29	21
STAT3	38	21
IL2 / IL-2R	36	18
TTR	19	17
VHL / HIF1-alpha	33	17
Transthyretin	16	16
BRD4 / NUT	16	15
BetaCatenin / Tcf4 and Tcf3	40	15
Rac1	15	15
Tubulin	14	14
SPIN1 / H3	15	10
CIAP1-BIR3/CASPASE-9	10	9
E2 / E1	11	9
MLLT1 / H3	16	8

DCN1/UBC12	7	7
PCNA trimer	7	7
c-Myc / Max	8	7
SOD1	10	5
RUNX1 / CBFb	4	4
TNFa / TNFa	5	4
VEGF / VEGFR	4	4
ZipA / ftsZ	4	4
53BP1 / H4	3	3
WD40 / H3	3	3
CRM1 / Rev	2	2
NRP / VEGF	2	2
UPAR / UPA	3	2
BRI1	1	1
CD4 / gp120	1	1
CD40 / CD154	1	1
CaM / CaMBD2	1	1
Rad51	1	1
TNFR1A / TNFB	1	1
Tak1 / Tab1	1	1

FIGURES

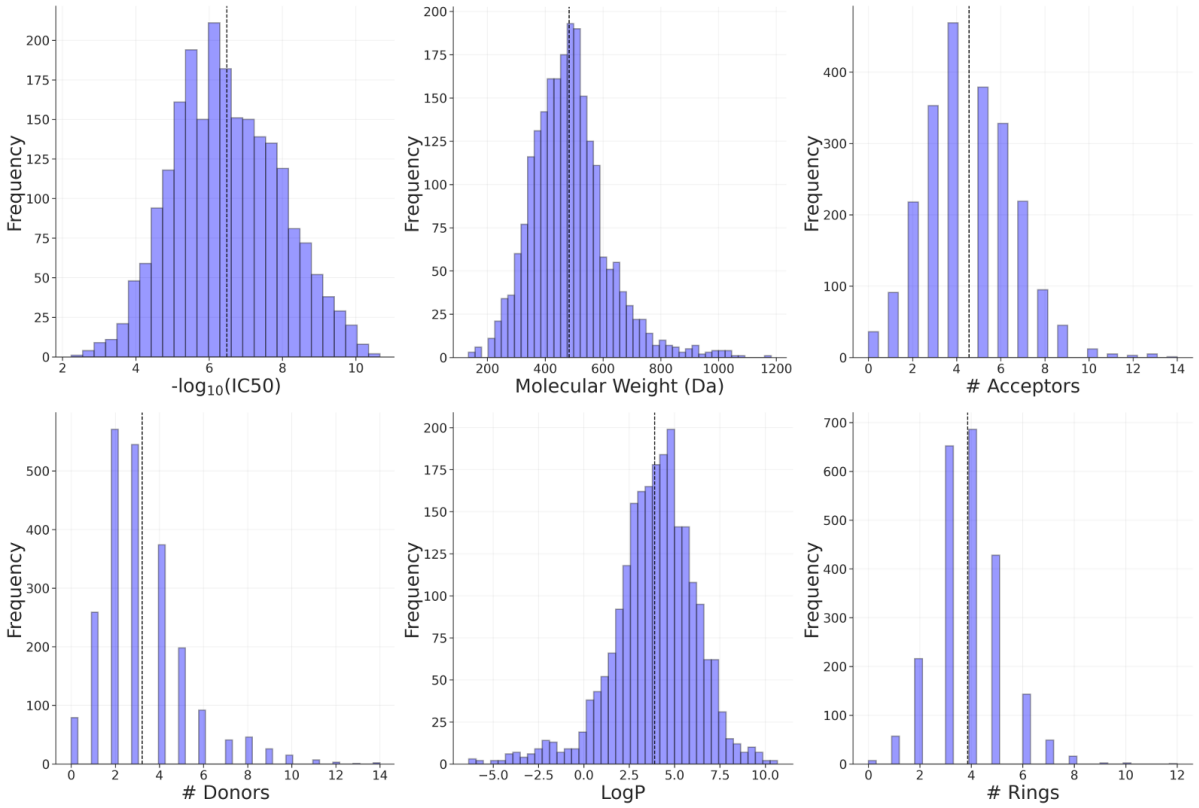


Figure S1 - Distribution of experimental IC50 values and property distribution of PPI inhibitors. The top-left histogram shows the distribution of experimental IC50 values for all inhibitors in the dataset after clustering with Tanimoto similarity of 0.8, , in terms of -$\log_{10}$(Molar). The remaining histograms depict the distribution of common physicochemical properties of the compounds, including molecular weight (in Da), logP, number of hydrogen acceptors and donors, and number of rings.

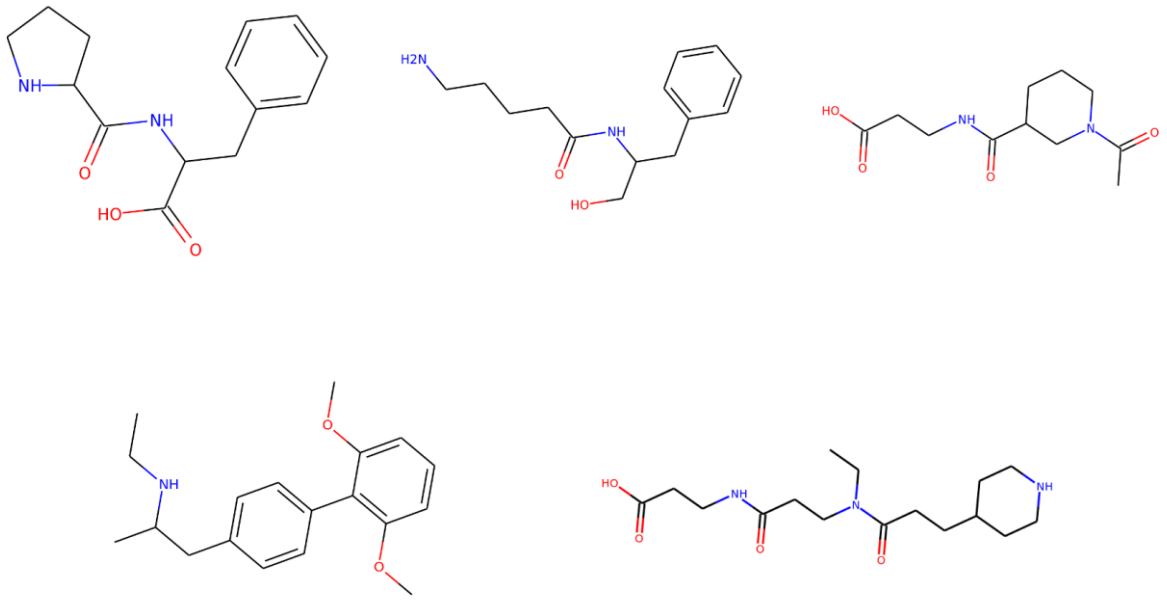


Figure S2. Frequent substructures within the dataset of PPI inhibitors. More potent (IC50 < 1uM) compounds are enriched with ring substructures, including biphenyls.

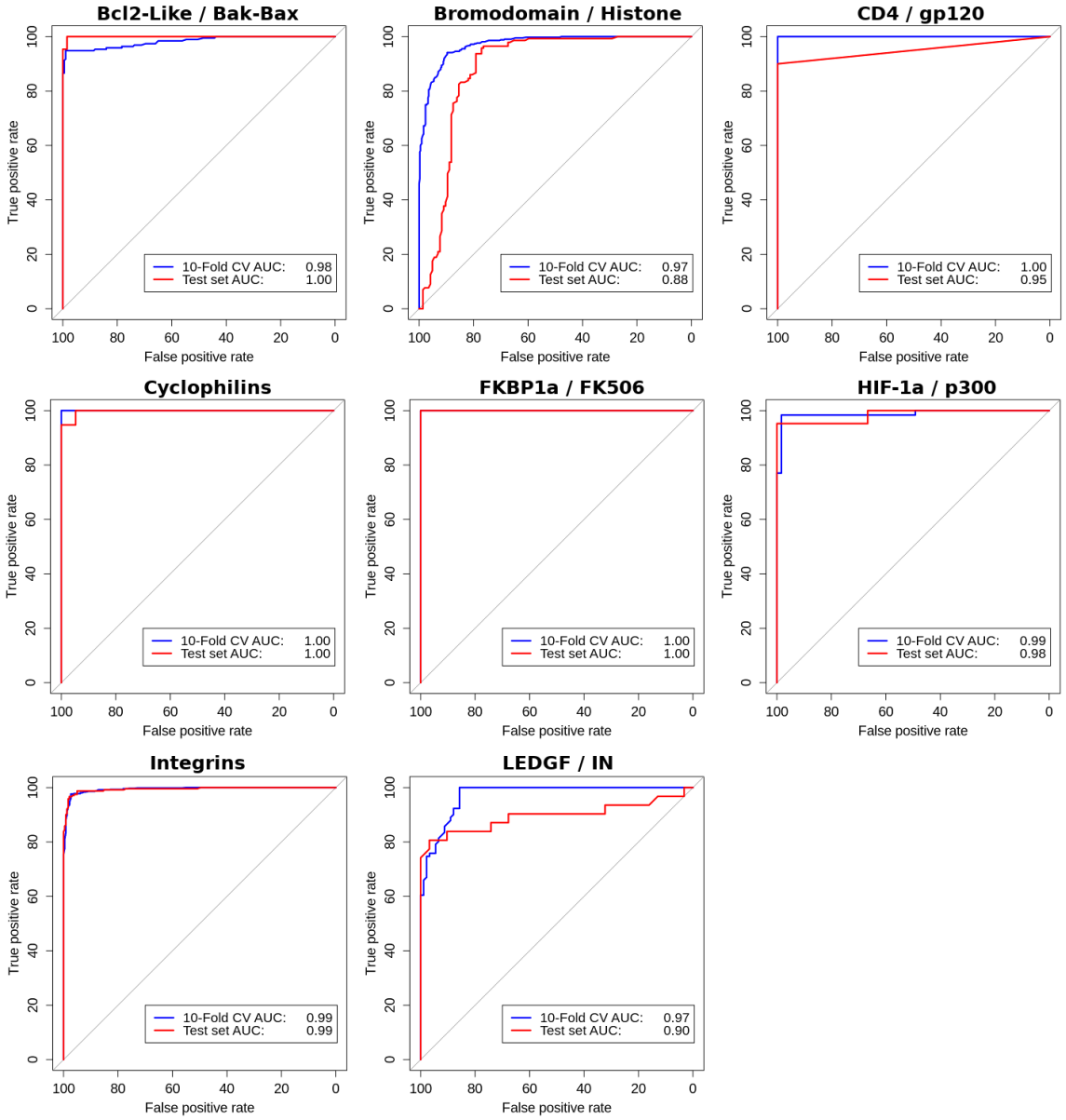


Figure S3. ROC curves for class-specific predictors during training under 10-Fold cross-validation and non-redundant test sets.

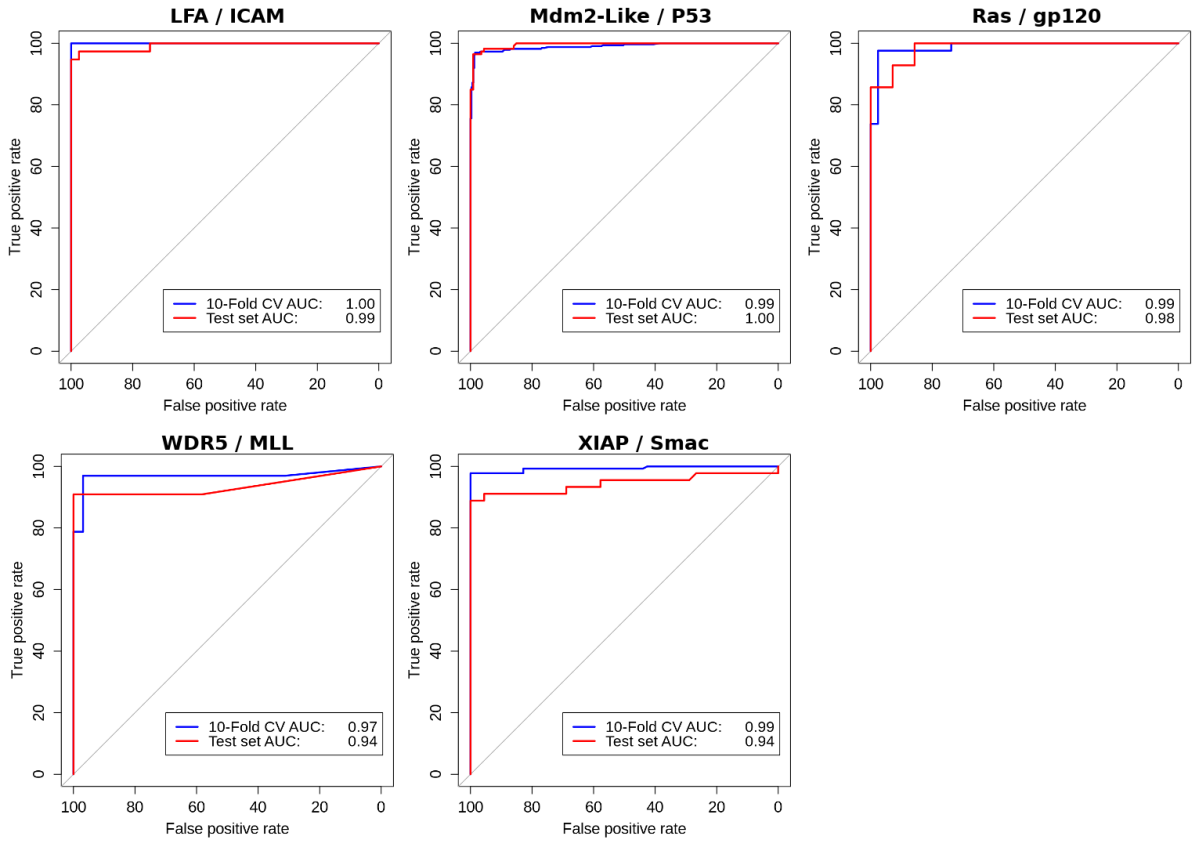


Figure S4. ROC curves for class-specific predictors on 10-Fold cross-validation and non-redundant test sets.

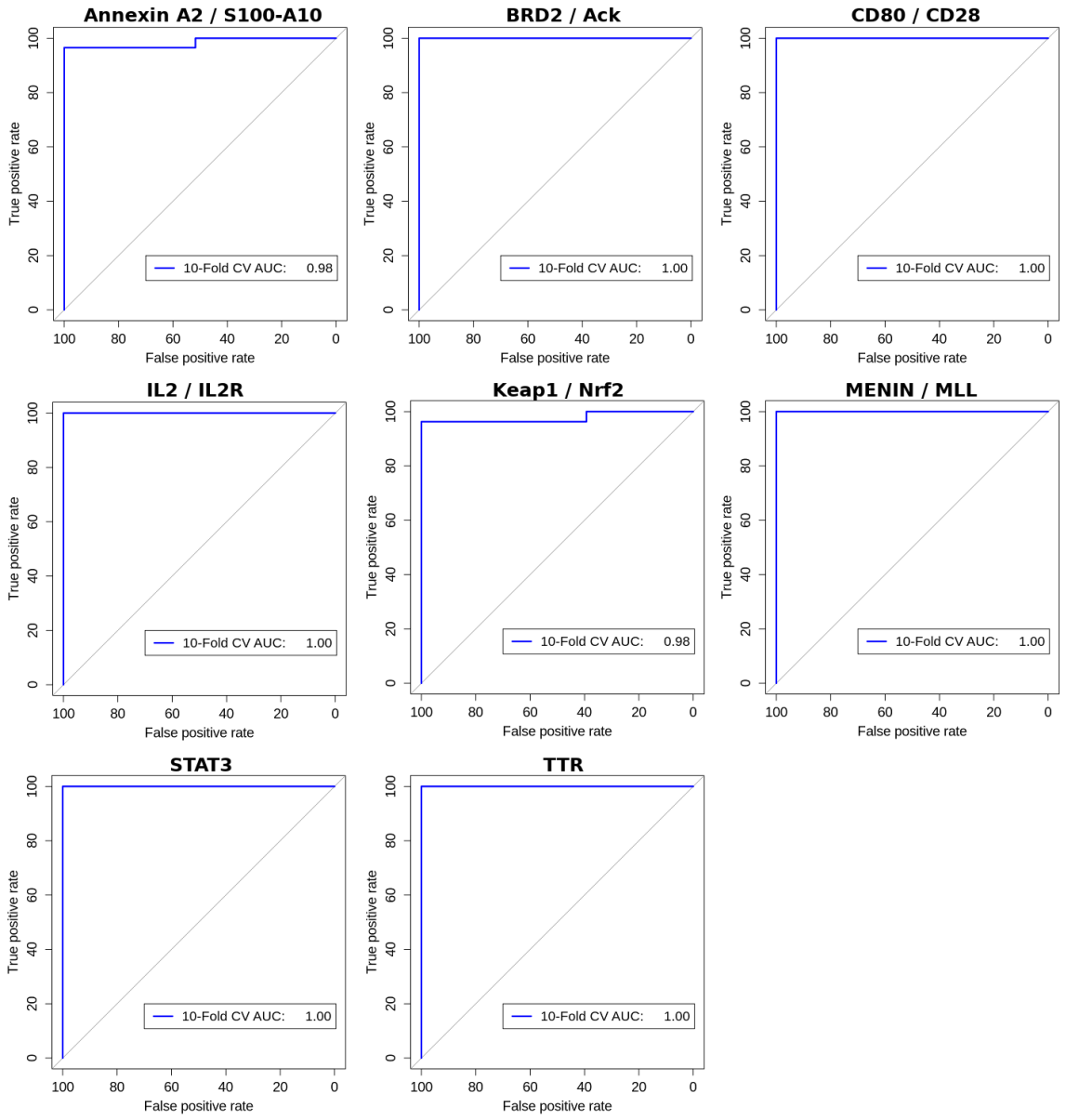


Figure S5. ROC curves for class-specific predictors with a limited number of inhibitors. Given the lack of data available for 8 PPI targets results are shown under 10-Fold cross-validation.

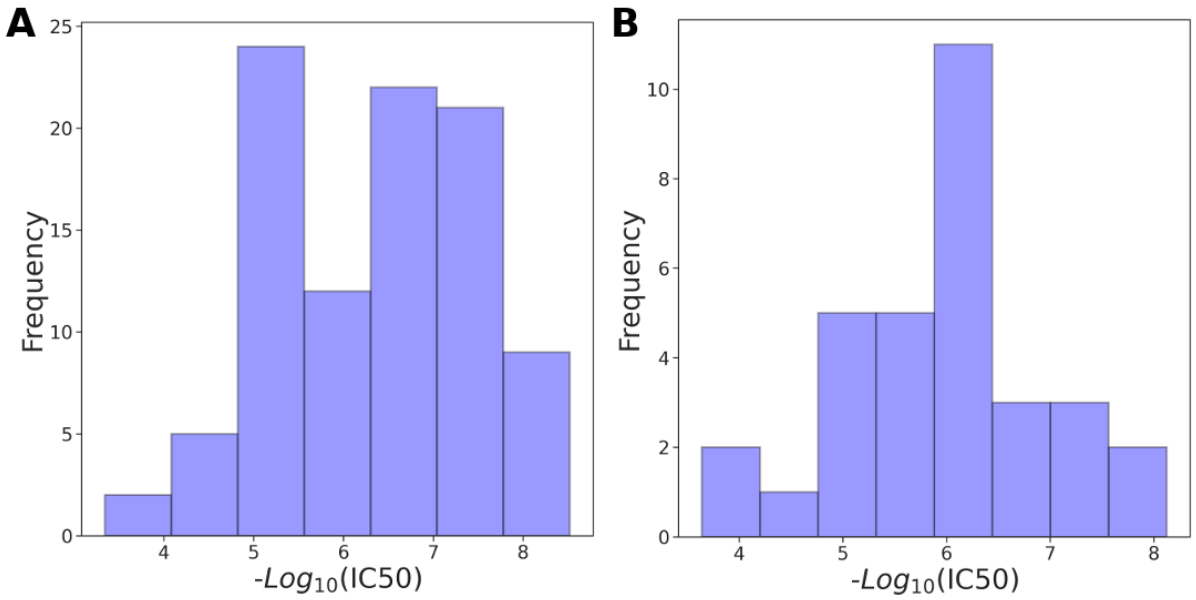


Figure S6. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the BCL2-Like / BAX-BAK complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

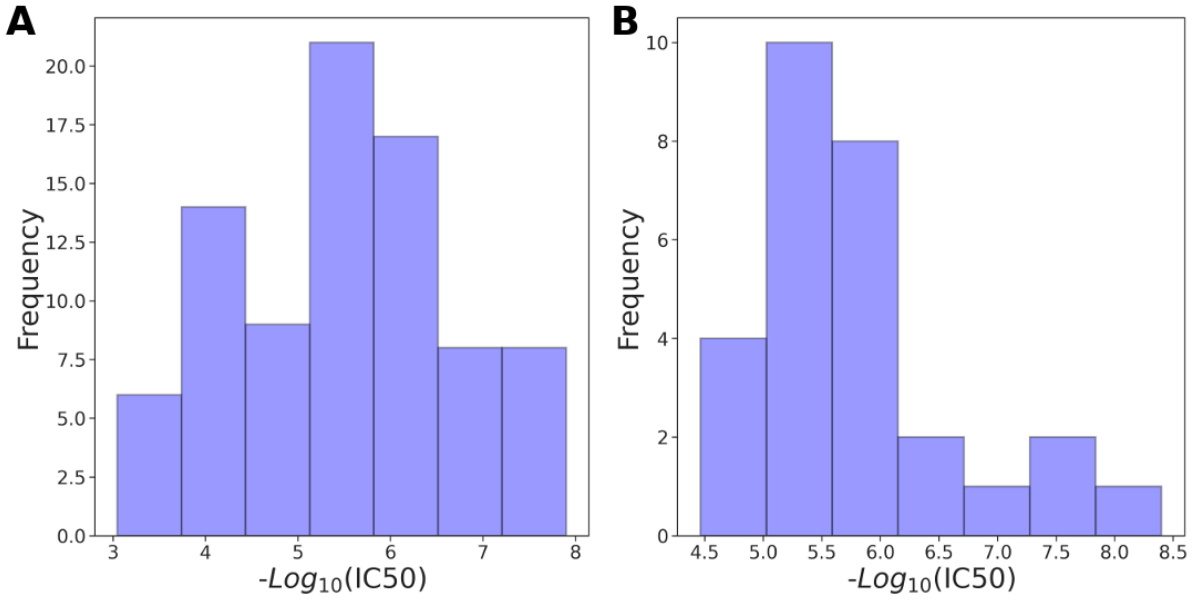


Figure S7. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the Bromodomain / Histone complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

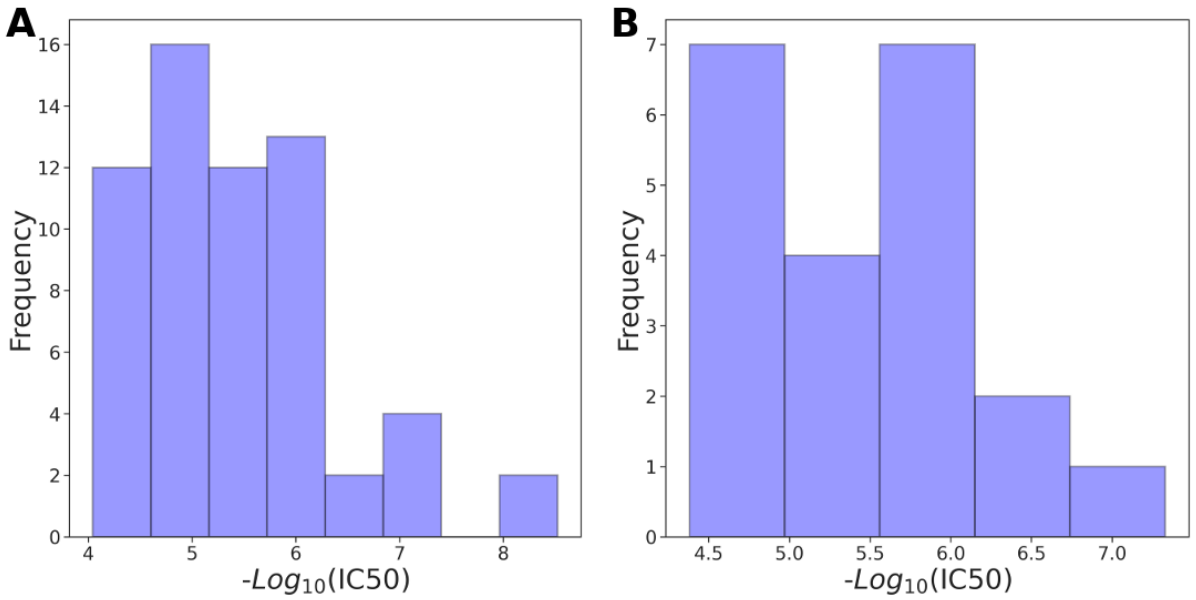


Figure S8. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the HIF-1a / p300 complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

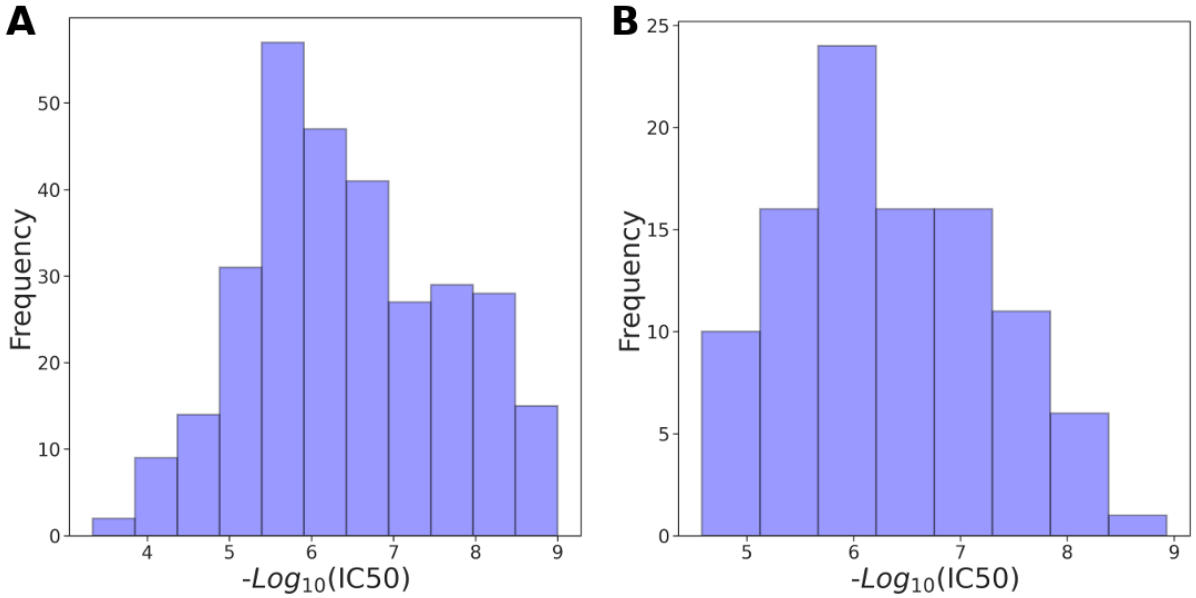


Figure S9. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the Mdm2-Like / P53 complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

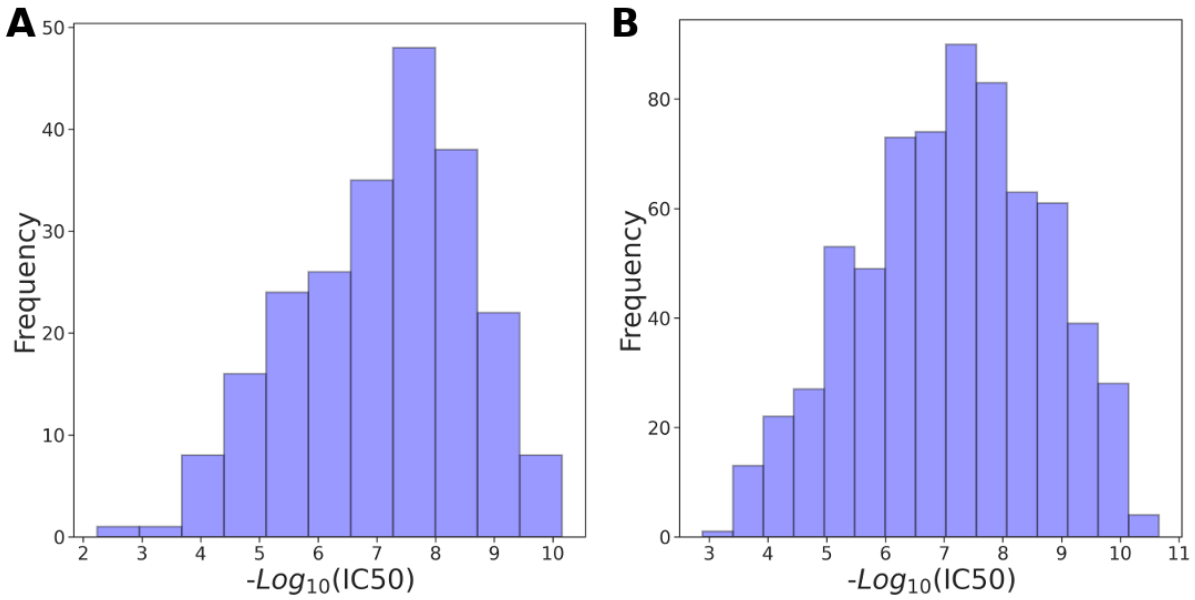


Figure S10. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the Integrins complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

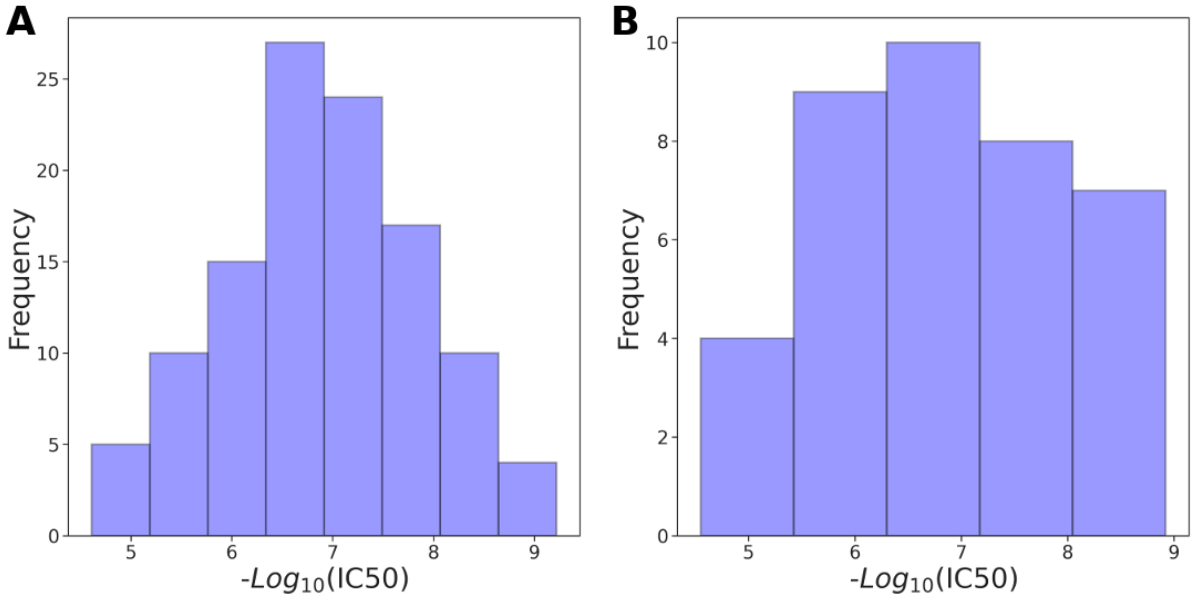


Figure S11. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the LFA / ICAM complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

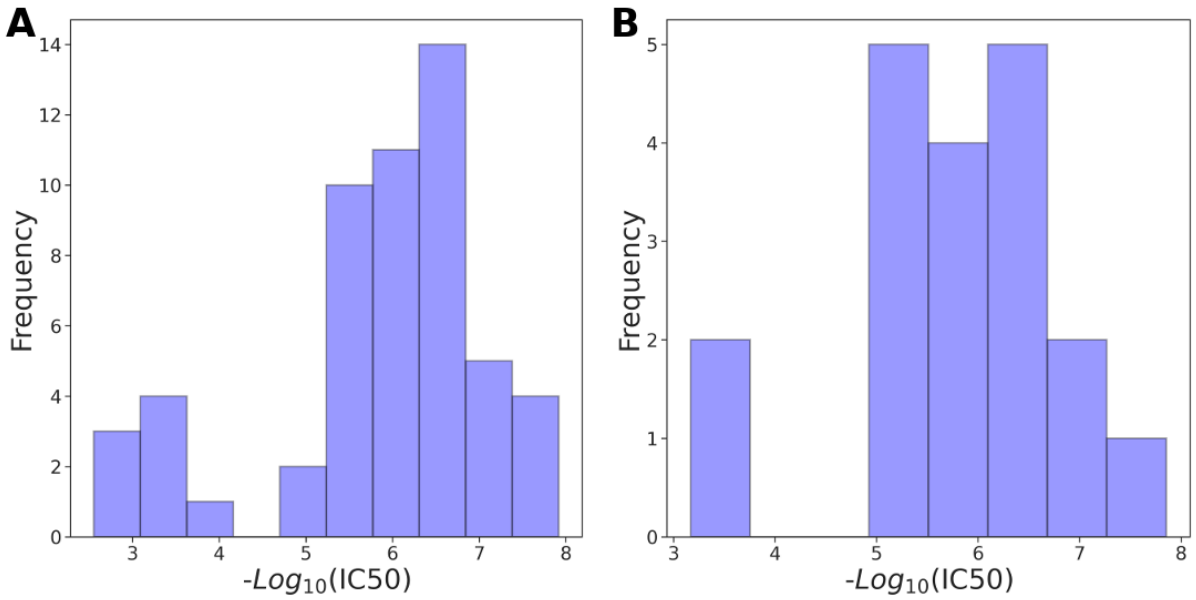


Figure S12. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the Cyclophilins complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

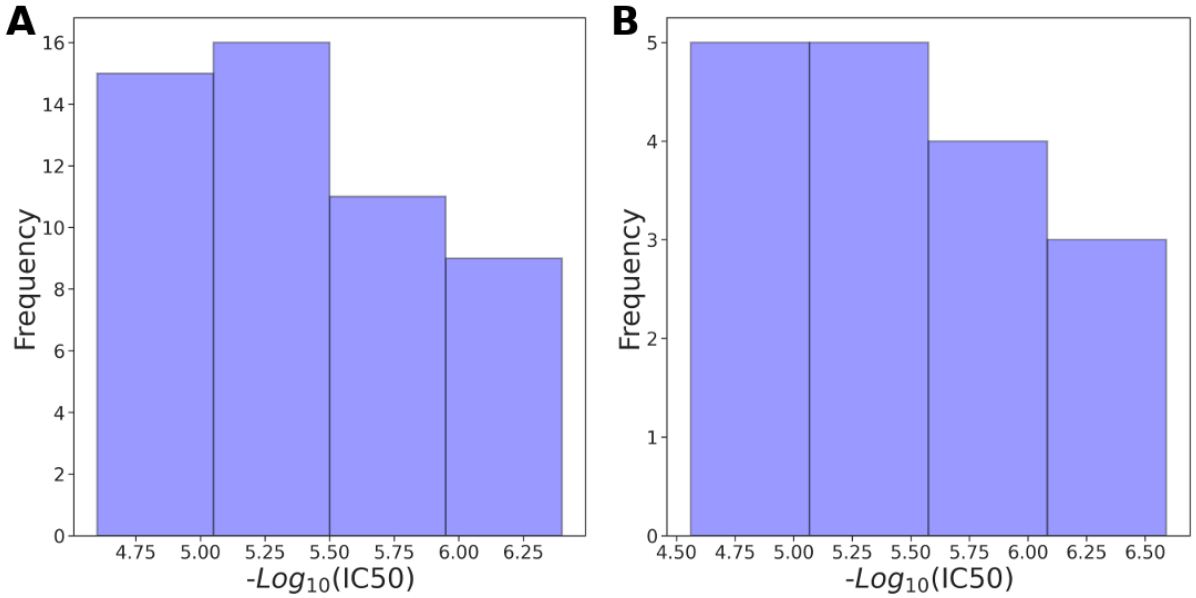


Figure S13. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the LEDGF / IN complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

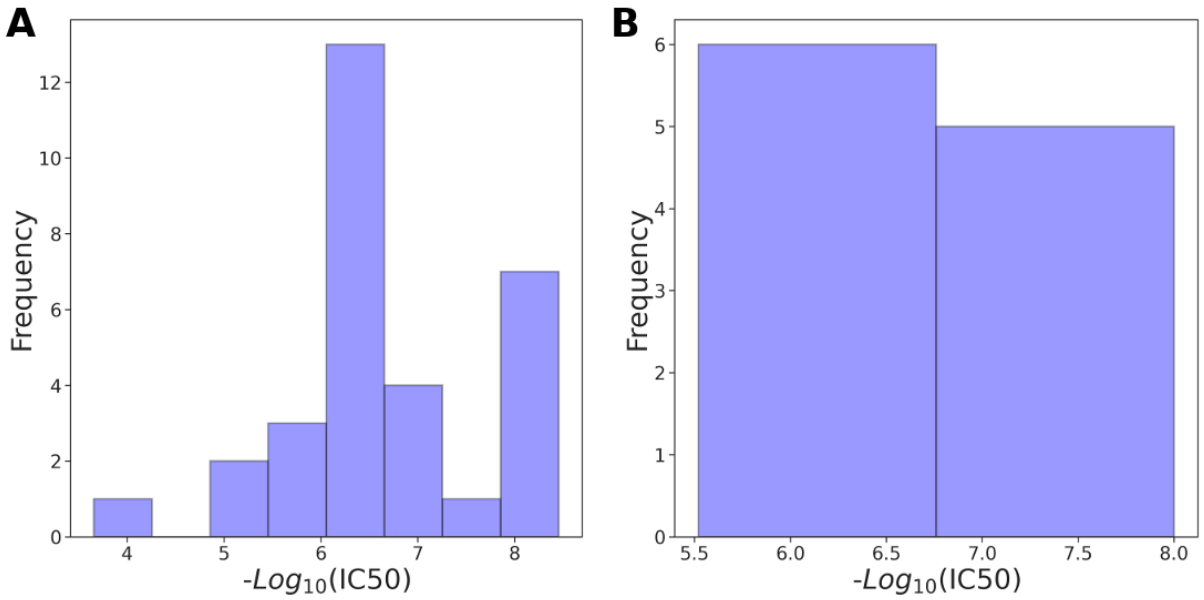


Figure S14. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors targeting the XIAP / Smac complex. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

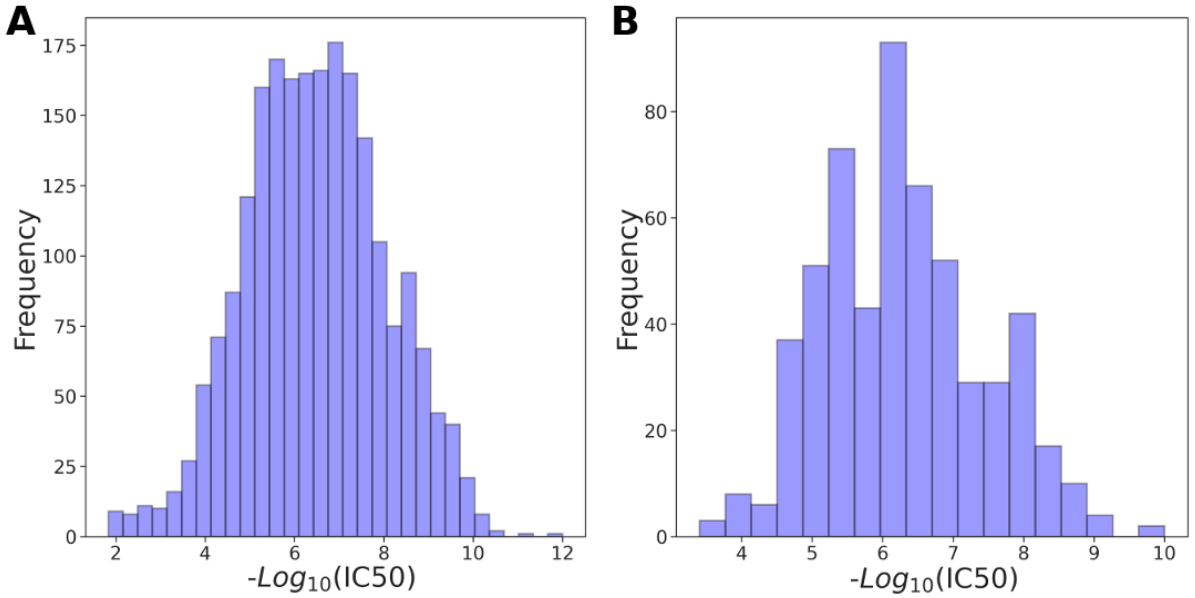


Figure S15. Distribution of PPI half maximal inhibitory concentration (IC50) for inhibitors used to build the general predictor of PPI inhibitory activity. A) depicts distribution on the set of molecules used for training the regression model and B) on the non-redundant test set. IC50 values are shown as -log10.

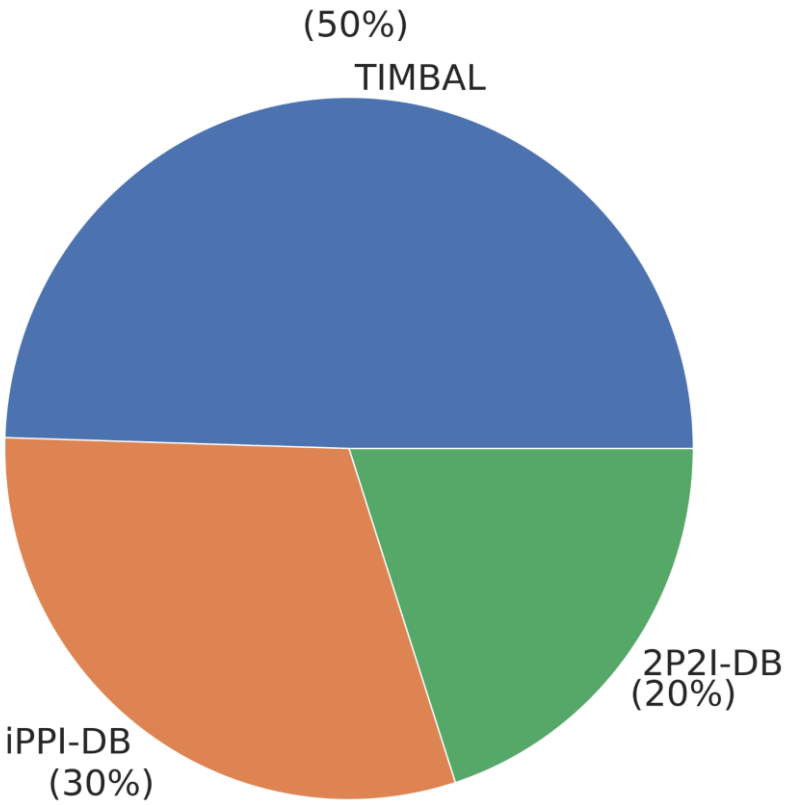
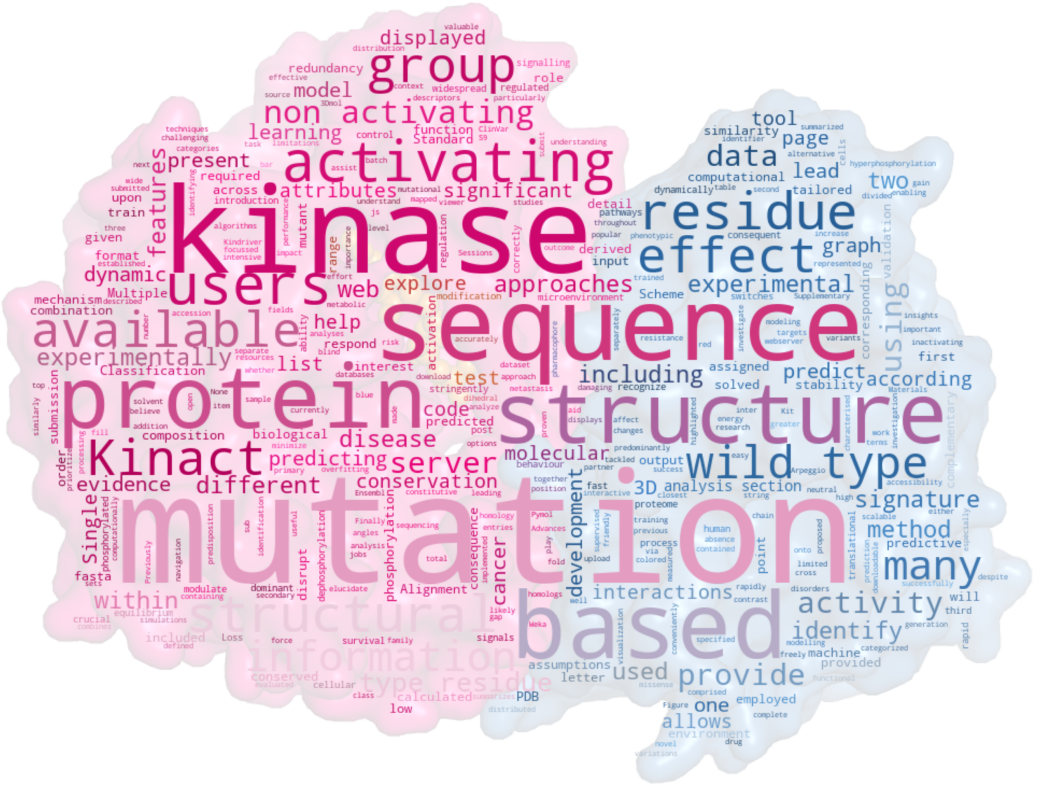


Figure S16. Distribution of PPI inhibitors retrieved from 3 different databases: TIMBAL, iPPI-DB and 2P2I-DB.

Chapter 6

Effects of mutations on phosphorylation mediated interactions



Published online 21 May 2018 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* ***W127–W132***
doi: 10.1093/nar/gky375

Kinact: a computational approach for predicting activating missense mutations in protein kinases

Carlos H.M. Rodrigues[1], David B. Ascher[1,2,3,*] and Douglas E.V. Pires[3,*]

[1]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, [2]Department of Biochemistry, University of Cambridge and [3]Instituto René Rachou, Fundação Oswaldo Cruz

Received January 31, 2018; Revised April 15, 2018; Editorial Decision April 26, 2018; Accepted April 28, 2018

ABSTRACT

Protein phosphorylation is tightly regulated due to its vital role in many cellular processes. While gain of function mutations leading to constitutive activation of protein kinases are known to be driver events of many cancers, the identification of these mutations has proven challenging. Here we present Kinact, a novel machine learning approach for predicting kinase activating missense mutations using information from sequence and structure. By adapting our graph-based signatures, Kinact represents both structural and sequence information, which are used as evidence to train predictive models. We show the combination of structural and sequence features significantly improved the overall accuracy compared to considering either primary or tertiary structure alone, highlighting their complementarity. Kinact achieved a precision of 87% and 94% and Area Under ROC Curve of 0.89 and 0.92 on 10-fold cross-validation, and on blind tests, respectively, outperforming well established tools ($P < 0.01$). We further show that Kinact performs equally well on homology models built using templates with sequence identity as low as 33%. Kinact is freely available as a user-friendly web server at http://biosig.unimelb.edu.au/kinact/.

INTRODUCTION

The ability of cells to recognize and correctly respond to their microenvironment is crucial for survival. In order to dynamically respond to cellular signals, fast dynamic switches are required. Protein phosphorylation is the most widespread type of post-translational modification, with over one-third of the proteins in the human proteome phosphorylated (1). The dynamic equilibrium between phosphorylation and dephosphorylation is stringently regulated, and provides a rapid mechanism to modulate protein behaviour and activity across most signalling pathways (2). Loss of control over this regulation process, through the introduction of dominant activating mutations in kinases and the consequent hyperphosphorylation of their targets can have many phenotypic consequences, including the development and metastasis of many cancers (3–7), and the development of other metabolic disorders (8).

Advances in next generation sequencing techniques are leading to the identification of a range of novel mutations, including in kinases. In the absence of experimental information, it is currently challenging to identify mutations that are likely to lead to constitutive activation of kinases. While many computational approaches have been proposed for predicting the effects of mutations that disrupt activity, these approaches have been shown to be of limited success to predict gain of function mutations, as also shown on this work, despite the important roles they play in many diseases, particularly in cancer.

To fill this gap, here we present Kinact, a machine learning-based predictive model and web server. Using our graph-based signatures, the method was tailored to accurately identify kinase activating mutations from a combination of sequence and structural information.

MATERIALS AND METHODS

Data sets

Mutations were derived from three mutational databases with experimental evidence of their functional consequence: Kindriver (9); ClinVar (10); and Ensembl (11). Kinase mutations were divided into two groups based upon the available experimental evidence: activating and non-activating mutations. The non-activating group is represented by variations that either disrupt activity (inactivating) or have no significant biological effect (neutral). The activating mutations were defined by a significant experimentally measured increase in kinase activity.

The complete data set contained 384 mutations (260 activating and 124 non-activating) distributed across 42 proteins, of which 256 (186 activating and 70 non-activating) could be mapped onto experimentally solved 3D structures. Supplementary Figures S1 and S2 of Supplementary Mate-

*To whom correspondence should be addressed. Email: douglas.pires@minas.fiocruz.br
Correspondence may also be addressed to David B. Ascher. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au

rials summarises the composition and the class distribution of mutations over the data set.

The dataset of mutations with experimental structures available, which account for 256 mutations, was randomly split into training and blind test sets. The proportion of activating and non-activating mutations on training and blind test sets is similar to observed on the original dataset as an attempt to prevent bias on the final method. The training set is comprised of 179 mutations (130 activating and 49 non-activating) that were used to train Kinact under 10-fold cross validation. The remaining 77 (56 activating and 21 non-activating) were used as blind test for validating the predictive model, minimizing the risk of overfitting. In order to assess the quality of the sub sets selected for training and blind test we repeated this process 20 times and the final version of the web server was built using the predictive model with best performance. Average and standard deviation values are reported on Supplementary Materials.

In addition, 41 mutations (24 activating and 17 non-activating in 14 kinases) that did not have experimentally solved structures available, therefore were not part of the original 256 mutations, had their structure modelled using homology modelling for further evaluation of Kinact predictive performance as a blind test.

Feature engineering

The task of predicting and understanding the effects of mutations in proteins at a molecular level has been tackled by approaches using different biological features, each with their own assumptions and limitations. Protein structural and sequence features have been the two most popular categories of attributes used by these computational methods. Sequence-based features have focussed predominantly on the analysis of sequence residue conservation throughout a protein family and homologs (12) and sequence composition (13). By contrast, previous studies have used a wide range of structural features, including secondary structure, solvent accessibility and dihedral angles (14,15). Significant effort has also been employed on more computationally intensive approaches to model mutation effects from the use of force fields and energy terms, to molecular dynamics simulations (16,17).

As an alternative, the use of graph-based structural signatures have been shown to be a scalable and effective approach for modeling the residue environment, which was successfully employed to train machine learning-based methods to predict and elucidate effects of mutations on protein stability and interactions with their partner (18–26). Moreover, these have also been used to provide insights into the molecular mechanisms of mutations and how they lead to disease and disease predisposition (27–33) and drug resistance (34–41). These graph-based signatures are predominantly composed of distance patterns extracted from the wildtype residue environment, which together with a pharmacophore modelling of its components, has been shown to be an effective way to model both geometry and physicochemical composition of protein regions.

Despite these diverse range of approaches, a combination of sequence and structural information has also been proven to be valuable when predicting damaging mutations (42,43). Based on these assumptions, graph-based signatures together with complementary sequence and structural information were used to build a predictive model. This complementary information included: (a) wild-type residue environment descriptors, (b) wild-type residue interactions, (c) predicted stability changes upon mutation, (d) sequence-based predicted effects on protein function and (e) the mCSM mutation pharmacophore modelling. A total of 82 different attributes (72 structural and 10 sequence-based) were calculated for each mutation in our dataset. These were then provided as evidence to train and test supervised learning algorithms using the Weka Tool Kit (44). The attributes used on this work were categorised into six different groups and summarised in Supplementary Table S1 of Supplementary Materials.

WEB SERVER

We have implemented Kinact as a user-friendly, freely available web server (http://biosig.unimelb.edu.au/kinact/). The server front end was built using Bootstrap framework version 3.3.7, while the back-end was built in Python via the Flask framework (Version 0.12.2). It is hosted on a Linux server running Apache.

Input

The server provides two different input options for the user (Supplementary Figure S4). The 'Single mutation' option allows users to predict whether a given mutation will lead to protein kinase activation or not. This option requires the user to provide a PDB (45) file or PDB accession code of the kinase, the point mutation specified as a string containing the wild-type residue one-letter code, its corresponding residue number and the mutant residue one-letter code, and the chain identifier of the wild-type residue. The primary sequence of the kinase of interest in fasta format is also required. The 'Mutation list' option allows users to upload a list of mutations in a file for batch processing. In order to aid users to submit their jobs, sample submission entries are available on the submission page and a help page is available via the top navigation bar.

Output

For the 'Single mutation' option, as shown in Figure 1, the web server displays in the output page the prediction outcome of Kinact, the details of the user input data, such as structure of wild-type and mutant residues, and also information on the kinase group in which the submitted structure was assigned to, based on sequence similarity according to the Standard Kinase Classification Scheme (46).

In addition, Kinact provides a set of analyses to help users investigate in greater detail the impact of the mutation. All resources displayed within the analysis section, including Pymol Sessions and the Multiple Sequence Alignment in fasta format, are made available for download.

The first item in the analysis section (Supplementary Figure S5) allows users to explore the 3D structure and the inter-residue interactions established by the wild-type residue, calculated by Arpeggio (47). Below this, users can

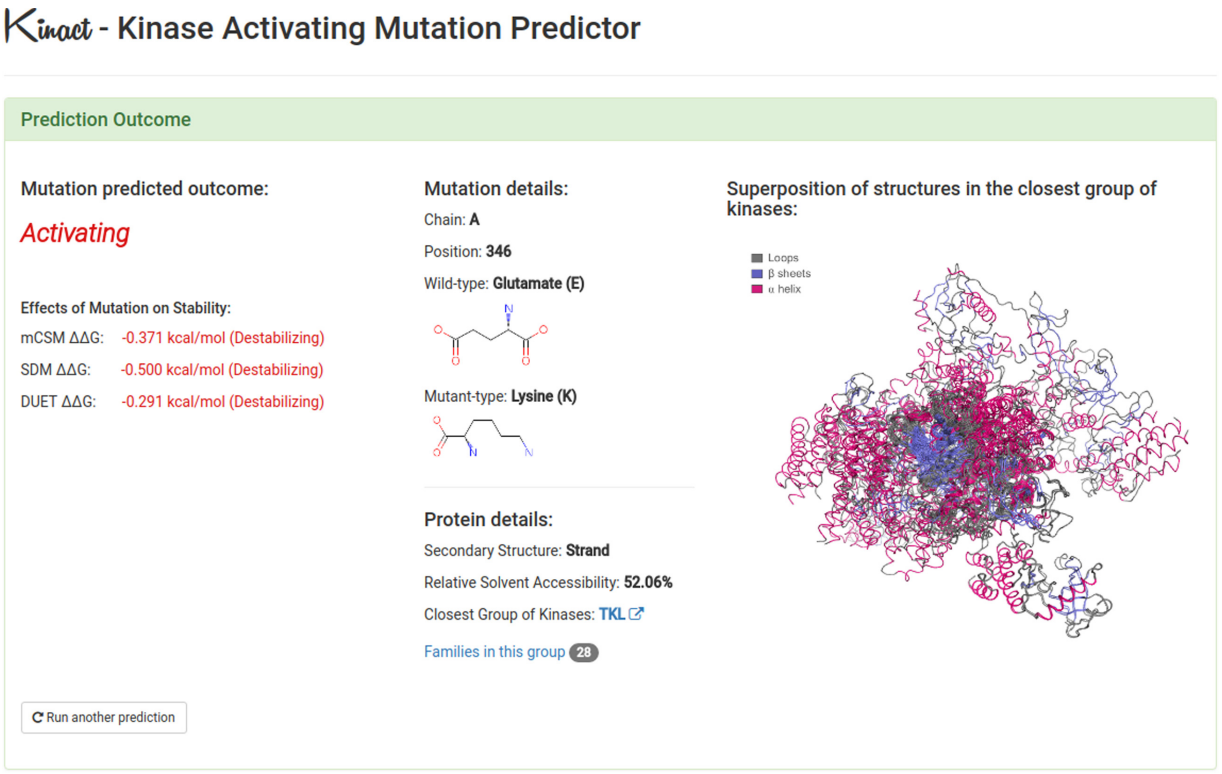


Figure 1. Web server results page for a single mutation prediction. The predicted outcome is shown alongside with complementary information on the submitted protein and the details of the mutation being evaluated. In addition, Kinact displays information on the group of homologue protein kinases according to the Standard Kinase Classification Scheme. The effects of mutation on protein stability calculated by mCSM (21), SDM (43) and DUET (25) are also shown.

also explore the conservation of residues with the structure of the wild-type kinase (Supplementary Figure S6). The 3D structure of the kinase of interest is displayed and colored according to conservation within the kinase sub-group, from red (not conserved) to blue (conserved). The structures are displayed in an interactive viewer implemented with 3Dmol.js (48).

Finally, users can also explore, within the analysis section, a multiple sequence alignment of the sequence of the provided structure and those from the closest kinase group according to the Standard Kinase Classification Scheme, assigned by similarity (Supplementary Figure S7). Previously experimentally characterised point mutations within any kinases of the group are highlighted, enabling users to rapidly identify through homology the effect of mutations at the corresponding residue position.

For the 'Mutation list' option, the server output is shown as a downloadable table (Supplementary Figure S8) and users also have the option to analyse each mutation separately, similarly to what was described for the 'Single mutation' option.

VALIDATION

In order to evaluate the quality of the training and blind test sets used we performed a resampling of these subsets 20 times and evaluated the performance of the predictive model on each split using AUC and precision. All values for the blind tests are reported on Supplementary Materials for each sample. Average and standard deviation are also shown and no bias was identified. Here we compare the performance of the best predictive model of Kinact with widely used tools to study the effects of mutations in proteins functions PolyPhen2 (42), SIFT (12) and wKinMut2 (49), a tool to identify and interpret pathogenic variants in human protein kinases.

Performance on cross validation

In order to better evaluate the contribution of structure and sequence-based attributes on the performance of supervised learning algorithms, three different predictive models were generated. The first model uses only attributes that rely on protein sequence information, which include mutation tolerance predictions (12,42), as well as a pharmacophore difference vector between wild-type and mutant residues, as proposed by the mCSM signatures (21), for this model we used the complete original dataset of 384 mutations. The second model uses only structural attributes calculated using the experimental structural data from the PDB. These include the graph-based structural signatures and complementary descriptors described in Supplementary Table S1 of Supplementary materials. Finally, the third model was constructed based on a combination of all attributes, using both sequence and structural data. For the models that used structural data on their predictions we used only the

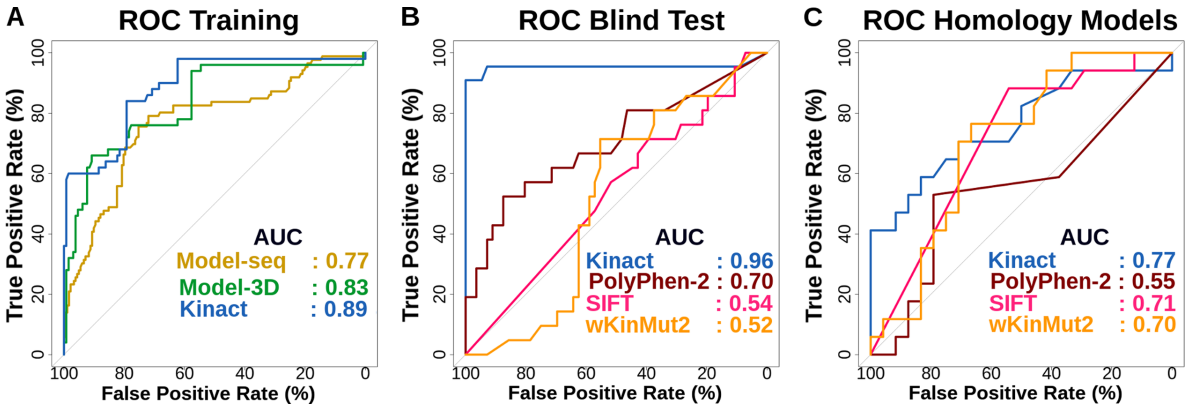


Figure 2. Comparative performance of Kinact. The ROC curves obtained for the training data set for models using sequence information alone, structural information alone, and the Kinact combined model is shown in (**A**). Kinact (AUC of 0.89), performs significantly better (*P*-value < 0.01) than the models using either just sequence or structural data (AUC of 0.77 and 0.83, respectively). In order to compare the performance of Kinact against the widely used tools SIFT, PolyPhen-2 and wKinMut2, a blind test (**B**) over a non-redundant test was evaluated and Kinact (AUC of 0.96) significantly (*P*-value < 0.01) outperformed all three methods (AUC of 0.54, 0.70 and 0.52, respectively). Using homology models (**C**), Kinact was also able to accurately identify activating mutations (AUC of 0.77), and again outperformed the other methods.

dataset of mutations with experimental structure available, which accounts for 256 mutations.

In order to run and assess the performance of the machine learning algorithms, we split each dataset into 70% of the mutations for training and 30% for blind test. In that sense, for the model that uses only sequence-based data we used 268 mutations for training (182 activating and 86 non-activating) and 116 for blind test (77 activating and 39 non-activating). For the other two models that used structure-based features 179 mutations were used for training and 77 mutations for blind test as previously described. All models were trained under 10-fold cross validation. Supplementary Figure S3 of Supplementary Materials summarises the distribution of activating and non-activating mutations in training and blind test sets for all models. Machine learning methods, evaluation procedures and performance metrics used are described in Supplementary Data.

A series of experiments were carried out to assess the performance of Kinact to predict whether a given mutation was likely to lead to constitutive activation of a kinase. The ROC curves across the training data set for models using sequence information alone, structure-based features alone, and the Kinact model that combines both attribute classes are shown in Figure 2. Details on the evaluation metrics for each algorithm are summarised on Supplementary Tables S2-S4 in Supplementary materials. Across the complete training set, Kinact achieved a Precision of 87% and Area Under ROC Curve of 0.89, significantly higher than the models using either just sequence or structural data (AUC of 0.77 and 0.83, and Precision of 0.78 and 0.81, respectively, $P < 0.01$). The final predictive models were trained using the full training set and all the performance evaluation metrics were calculated considering the average values for all 10 folds from cross validation.

Blind test

In order to properly evaluate the method's predictive performance and generalization, Kinact was initially evaluated against a separate, independent, non-redundant blind test set comprised of 77 missense mutations in protein kinases with available experimental structures, achieving a precision of 97% and Area Under ROC Curve of 0.96. When comparing with other methods, Kinact significantly outperformed (Figure 2B) all three methods (*P*-value < 0.01). Looking specifically at the activating mutations, SIFT predicted 55% of mutations as deleterious (score < 0.05), PolyPhen-2 classified 84% as probably damaging (score > 0.85), and wKinMut2 predicted 62% of mutations as disease related (score > 0), while Kinact correctly classified 99% of them. Comparisons of Kinact with tools that assess the effects of mutations on protein stability are described on Supplementary Materials.

Homology models

The performance of the web server to accurately classifying mutations using homology models was evaluated using a set of 41 mutations in kinases without experimentally resolved structures. Homology models of the kinases were generated by Modeller (50) using experimentally resolved structures down to 33% sequence identity. Using the homology models, Kinact was able to accurately identify activating mutations (AUC of 0.77 and precision of 0.78), providing confidence and robustness in the applicability of this approach beyond experimental structures to those that are computationally modelled. This was also significantly better than PolyPhen-2, SIFT and wKinMut2 (Figure 2C). When comparing the performance of the methods specifically at the activating mutations, Kinact was able to classify correctly 100% of mutations, while SIFT predicted 75% as deleterious (score < 0.05), PolyPhen-2 classified 83% as probably damaging (score > 0.85), and wKinMut2 predicted 77% as disease related (score > 0).

CONCLUSION

We present here, Kinact, a predictive model and web server tailored for identifying kinase activating mutations using

graph-based signatures, sequence and structural data. Kinact conveniently combines high-performance, open access, web visualization tools to assist research on how mutations affect protein kinases activity as well as prioritise mutations for further investigation. Given the importance of these variants in the context of many diseases, especially on the development of many types of cancer, and also that widely used tools have not been able to successfully predict gain of function mutations, we believe Kinact will be a useful tool to help identify and understand the role of these mutations. The method is freely available as a user friendly and easy to use web server at http://biosig.unimelb.edu.au/kinact/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Australian Government Research Training Program Scholarship [to C.H.M.R]; Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; Victorian Life Sciences Computation Initiative (VLSCI), an initiative of the Victorian Government, Australia, on its Facility hosted at the University of Melbourne [UOM0017]; Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P.]; Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.]. Funding for open access charge: Instituto René Rachou (IRR/FIOCRUZ Minas).
Conflict of interest statement. None declared.

REFERENCES

1. Cohen,P. (2002) The origins of protein phosphorylation. *Nat. Cell Biol.*, **4**, E127–E130.
2. Salazar,C. and Hofer,T. (2009) Multisite protein phosphorylation–from molecular mechanisms to kinetic models. *FEBS J.*, **276**, 3177–3198.
3. Bose,R., Kavuri,S.M., Searleman,A.C., Shen,W., Shen,D., Koboldt,D.C., Monsey,J., Goel,N., Aronson,A.B., Li,S. *et al.* (2013) Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discov.*, **3**, 224–237.
4. Chirgadze,D.Y., Ascher,D.B., Blundell,T.L. and Sibanda,B.L. (2017) DNA-PKcs, allostery, and DNA double-strand break repair: defining the structure and setting the stage. *Methods Enzymol.*, **592**, 145–157.
5. Grabiner,B.C., Nardi,V., Birsoy,K., Possemato,R., Shen,K., Sinha,S., Jordan,A., Beck,A.H. and Sabatini,D.M. (2014) A diverse array of cancer-associated MTOR mutations are hyperactivating and can predict rapamycin sensitivity. *Cancer Discov.*, **4**, 554–563.
6. Sibanda,B.L., Chirgadze,D.Y., Ascher,D.B. and Blundell,T.L. (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. *Science*, **355**, 520–524.
7. Tiacci,E., Pettirossi,V., Schiavoni,G. and Falini,B. (2017) Genomics of hairy cell leukemia. *J. Clin. Oncol.*, **35**, 1002–1010.
8. Lahiry,P., Torkamani,A., Schork,N.J. and Hegele,R.A. (2010) Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.*, **11**, 60–74.
9. Simonetti,F.L., Tornador,C., Nabau-Moreto,N., Molina-Vila,M.A. and Marino-Buslje,C. (2014) Kin-Driver: a database of driver mutations in protein kinases. *Database (Oxford)*, **2014**, bau104.
10. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
11. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
12. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
13. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
14. Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
15. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
16. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
17. Zimmermann,M.T., Urrutia,R., Oliver,G.R., Blackburn,P.R., Cousin,M.A., Bozeck,N.J. and Klee,E.W. (2017) Molecular modeling and molecular dynamic simulation of the effects of variants in the TGFBR2 kinase domain as a paradigm for interpretation of variants obtained by next generation sequencing. *PLoS One*, **12**, e0170822.
18. Pires,D.E. and Ascher,D.B. (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.*, **44**, W469–W473.
19. Pires,D.E. and Ascher,D.B. (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res.*, **44**, W557–W561.
20. Pires,D.E. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.
21. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
22. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res.*, **43**, D387–D391.
23. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.*, **6**, 29575.
24. Pires,D.E., Chen,J., Blundell,T.L. and Ascher,D.B. (2016) In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.*, **6**, 19848.
25. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–W319.
26. Rodrigues,C.H., Pires,D.E. and Ascher,D.B. (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.*, doi:10.1093/nar/gky300.
27. Casey,R.T., Ascher,D.B., Rattenberry,E., Izatt,L., Andrews,K.A., Simpson,H.L., Challis,B., Park,S.M., Bulusu,V.R., Lalloo,F. *et al.* (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol. Genet. Genomic Med.*, **5**, 237–250.
28. Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.
29. Nemethova,M., Radvanszky,J., Kadasi,L., Ascher,D.B., Pires,D.E., Blundell,T.L., Porfirio,B., Mannoni,A., Santucci,A., Milucci,L. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur. J. Hum. Genet.*, **24**, 66–72.
30. Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B.,

Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom. Med.*, **2**, 7.
31. Trezza,A., Bernini,A., Langella,A., Ascher,D.B., Pires,D.E.V., Sodi,A., Passerini,I., Pelo,E., Rizzo,S., Niccolai,N. *et al.* (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest. Ophthalmol. Vis. Sci.*, **58**, 5320–5328.
32. Usher,J.L., Ascher,D.B., Pires,D.E., Milan,A.M., Blundell,T.L. and Ranganath,L.R. (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: Identification of novel mutations. *JIMD Rep.*, **24**, 3–11.
33. Hnizda,A., Fabry,M., Moriyama,T., Pachl,P., Kugler,M., Brinsa,V., Ascher,D.B., Carroll,W.L., Novak,P., Zaliova,M. *et al.* (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. *Leukemia*, **In Press**.
34. Albanaz,A.T.S., Rodrigues,C.H.M., Pires,D.E.V. and Ascher,D.B. (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin. Drug Discov.*, **12**, 553–563.
35. Pandurangan,A.P., Ascher,D.B., Thomas,S.E. and Blundell,T.L. (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.*, **45**, 303–311.
36. Phelan,J., Coll,F., McNerney,R., Ascher,D.B., Pires,D.E., Furnham,N., Coeck,N., Hill-Cawthorne,G.A., Nair,M.B., Mallard,K. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.*, **14**, 31.
37. Hawkey,J., Ascher,D.B., Judd,L.M., Wick,R.R., Kostoulias,X., Cleland,H., Spelman,D.W., Padiglione,A., Peleg,A.Y. and Holt,K.E. (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. *Microb.Genomics*, **4**, e000165.
38. Vedithi,S.C., Malhotra,S., Das,M., Daniel,S., Kishore,N., George,A., Arumugam,S., Rajan,L., Ebenezer,M., Ascher,D.B. *et al.* (2018) Structural implications of mutations conferring rifampin resistance in mycobacterium leprae. *Sci. Rep.*, **8**, 5016.
39. Karmakar,M., Globan,M., Fyfe,J.A.M., Stinear,T.P., Johnson,P.D.R., Holmes,N.E., Denholm,J.T. and Ascher,D.B. (2018) Analysis of a novel pncA mutation for susceptibility to Pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.*, **In Press**.
40. Singh,V., Donini,S., Pacitto,A., Sala,C., Hartkoorn,R.C., Dhar,N., Keri,G., Ascher,D.B., Mondesert,G., Vocat,A. *et al.* (2017) The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. *ACS Infect. Dis.*, **3**, 5–17.
41. Holt,K.E., McAdam,P., Thai,P.V.K., Thuong,N.T.T., Ha,D.T.M.H., Lan,N.N., Lan,N.H., Nhu,N.T.Q., Hai,H.T., Ha,V.T.N. *et al.* (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for EsxW Beijing variant in Vietnam. *Nat. Genet.*, **In Press**.
42. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
43. Pandurangan,A.P., Ochoa-Montano,B., Ascher,D.B. and Blundell,T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.*, **45**, W229–W235.
44. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H. (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.
45. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
46. Manning,G., Whyte,D.B., Martinez,R., Hunter,T. and Sudarsanam,S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
47. Jubb,H.C., Higueruelo,A.P., Ochoa-Montano,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol*, **429**, 365–371.
48. Rego,N. and Koes,D. (2015) 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, **31**, 1322–1324.
49. Vazquez,M., Pons,T., Brunak,S., Valencia,A. and Izarzugaza,J.M. (2016) wKinMut-2: identification and Interpretation of Pathogenic Variants in Human Protein Kinases. *Hum. Mutat.*, **37**, 36–42.
50. Webb,B. and Sali,A. (2014) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **1137**, 1–15.

SUPPLEMENTARY MATERIAL

Kinact: a computational approach for predicting activating missense mutations in protein kinases

Carlos H.M. Rodrigues[1], David B. Ascher[1,2,3,*], Douglas E.V. Pires[3,*]

[1]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne;
[2]Department of Biochemistry, University of Cambridge;
[3]Instituto René Rachou, Fundação Oswaldo Cruz

*To whom correspondence should be addressed. D.E.V.P. douglas.pires@minas.fiocruz.br; Correspondence may also be addressed to D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au or da382@cam.ac.uk

EVALUATION METRICS

A set of well-established and widely used performance metrics for classification algorithms were used to evaluate Kinact on both 10-fold cross validation and on blind tests. These metrics include Area Under ROC Curve (AUC), Precision, Recall and F-Measure. Such measurements are expressed based on the values of a binary contingency table, also known as a confusion matrix (Figure S1), where the classes are represented by convention as + (positive) and - (negative) signs.

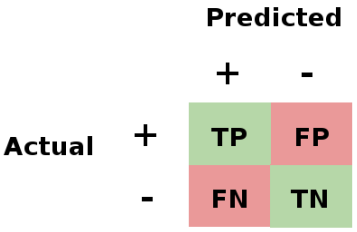


Figure 1 - Confusion matrix (actual *vs.* predicted). True and False Positives (TP and FP) indicate the number of predicted positives that were correctly and incorrectly classified, respectively. Similarly, True and False Negatives (TN and FN) refer to correct and wrong predictions for the negative class. The sum TP+FP+TN+FN is equal to the total amount number of instances in the data set being used.

Area Under ROC Curve (AUC)

The measure of Area Under the ROC Curve (AUC or AUROC) considers the True Positive Rate (TPR), also known as sensitivity, that corresponds to the proportion of positive data points that are correctly considered as positive; and also the False Positive Rate (FPR) that corresponds to the proportion of negative data that are wrongly considered as positive, regarding all negative data points. A Receiver Operating Curve (ROC) is then plotted using TPR versus FPR and the AUC is the area under such curve (1). Like precision, recall and f-measure, AUC values range from 0 to 1, which the later denoting a perfect classifier. A random binary classifier would generate an AUC of 0.5.

Precision

Precision denotes the proportion of Predicted Positive cases that are Actual Positives. It is defined by TP/(TP+FP).

Recall

Recall is defined as the proportion of Predicted Positives cases that are Actual Positives over all Predicted Positives. Using the convention defined in Figure S1, it is defined as TP/(TP+FN).

F-Measure

F-measure is the harmonic mean between Precision and Recall as defined by the formula below.

$$2\,x\,\frac{Precision\,x\,Recall}{Precision + Recall}$$

MACHINE LEARNING ALGORITHMS

Several classification algorithms from different paradigms implemented on the Weka Tool Kit (2) were considered during training.

Random Forest

The Random Forest algorithm uses a combination of decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the “forest”. The generalization error for forests converge to a limit as the number of trees in the forest become large (3). It is a fast and relatively easy to implement algorithm, produce highly accurate predictions and can handle a very large number of input variables without overfitting, given that all the trees are built from scratch without any previous information on the other trees in the forest and also the final prediction is the average of all the predictions for each tree. In fact, it is considered to be on of the most accurate general-purpose learning techniques available.

Classification by Regression with M5P

Classification via regression handles the discrete classes (nominal) of the data set as continuous labels (probability) in a probabilistic classification manner (4). The classification is achieved by defining a threshold, for example a prediction with a probability $\hat{y} < 0.5$ indicates non-activating and consequently $\hat{y} >= 0.5$ results in

activating output prediction, also known as linear decision boundary. Thus, algorithms that use this type of classification seeks for a model that generates the greatest approximate probability function that separates the classes in the dataset. In this sense, for the scope of this work, we used the Decision Trees M5P (5).

MLP

Multi-Layer Perceptron, also known as MLP, is a feed-forward neural network, consisting of many units, called neurons, which are connected by weighted links. The units are organised in several layers, namely an input layer, one or more hidden layers, and an output layer. The input layer receives an external activation vector, and forwards it via weighted connections to the units in the first hidden layer. These compute their activations and pass them to neurons in succeeding layers. From a distal point of view, an arbitrary input vector is propagated forward through the network, finally causing an activation vector in the output layer (6). The entire network function, that maps the input vector onto the output vector is determined by the connection weights of the network.

Decision Tree - J48

A decision tree is an algorithm that simulates trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on the feature values. The basic assumption made in the decision trees is that instances with different classes have different values in at least one of their features. One of the most useful characteristics of such algorithm is their comprehensibility. One can easily understand why the algorithm classifies an instance as belonging to a specific class by just looking at the generated tree and analyzing its rules (7). In this sense, the J48 algorithm is a variation of the C4.5 and ID3 algorithms (8) that uses information theory principles to evaluate how “good” an instance is, in the sense that it chooses the test that extract the maximum amount of information from a given set of cases.

QUALITY ASSESSMENT OF TRAINING AND BLIND TEST SETS

In order to evaluate the quality of the training and blind test sets we performed a resampling of these subsets 20 times and evaluated the performance of the predictive model on each split using AUC and precision. All values for the blind tests are reported below with averages and standard deviations at the bottom of the table. The split with best performance, which was the one used to build the final version of the server is highlighted.

TABLE 1 – Performance evaluation for 20 resamplings of training and blind sets. AUC and Precision for the blind test set are shown for each split.

Split	AUC	Precision
1	0.91	0.93
2	0.94	0.94
3	0.89	0.92
4	0.90	0.91
5	0.91	0.93
6	0.90	0.92
7	0.92	0.92
8	0.89	0.97
9	0.93	0.94
10	0.92	0.94
11	***0.96***	***0.97***
12	0.90	0.93
13	0.89	0.91
14	0.95	0.97
15	0.93	0.94
16	0.93	0.94
17	0.89	0.92
18	0.91	0.93
19	0.95	0.96
20	0.90	0.91
Average	*0.92*	*0.94*
Standard Deviation (σ)	*0.02*	*0.02*

COMPARISON WITH METHODS THAT ASSESS THE EFFECTS OF MUTATIONS ON PROTEIN STABILITY

In order to understand the relationship between the scores for predicted effects of mutations on stability given by mCSM, SDM, DUET and I-Mutant2 and the validation status of the mutations on protein kinases in our dataset (activating and non-activating), also comparing their predictive performance with Kinact, the distributions of all four scores for each status separately were analysed. We performed t-test with a 95% confidence interval to evaluate whether a significant difference between the scores for the two classes existed. No significant difference was observed. These analyses were summarised on the figure below.

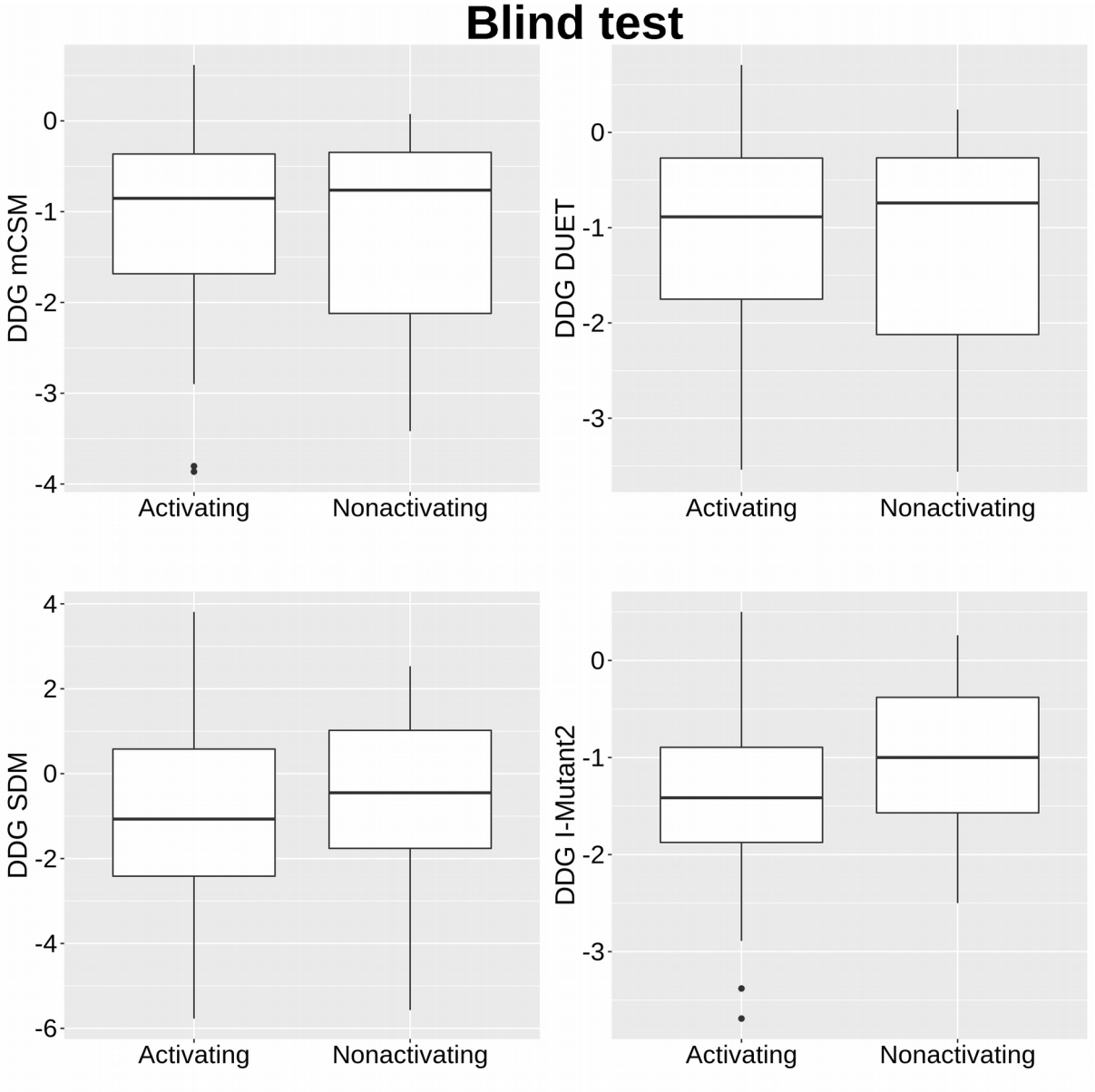


Figure 2 – Distribution of values for mCSM, SDM, DUET and I-Mutant2 for each validation status (Activating and Non-activating) on the blind test used to validate Kinact. T-test with 95% confidence interval was performance and no significant difference was observed for none of the scores.

A similar analysis was performed for the subset of mutations on structures modelled by homology modelling and again no statistical difference was observed. The distributions for each score is represented on the boxplots on the figure below.

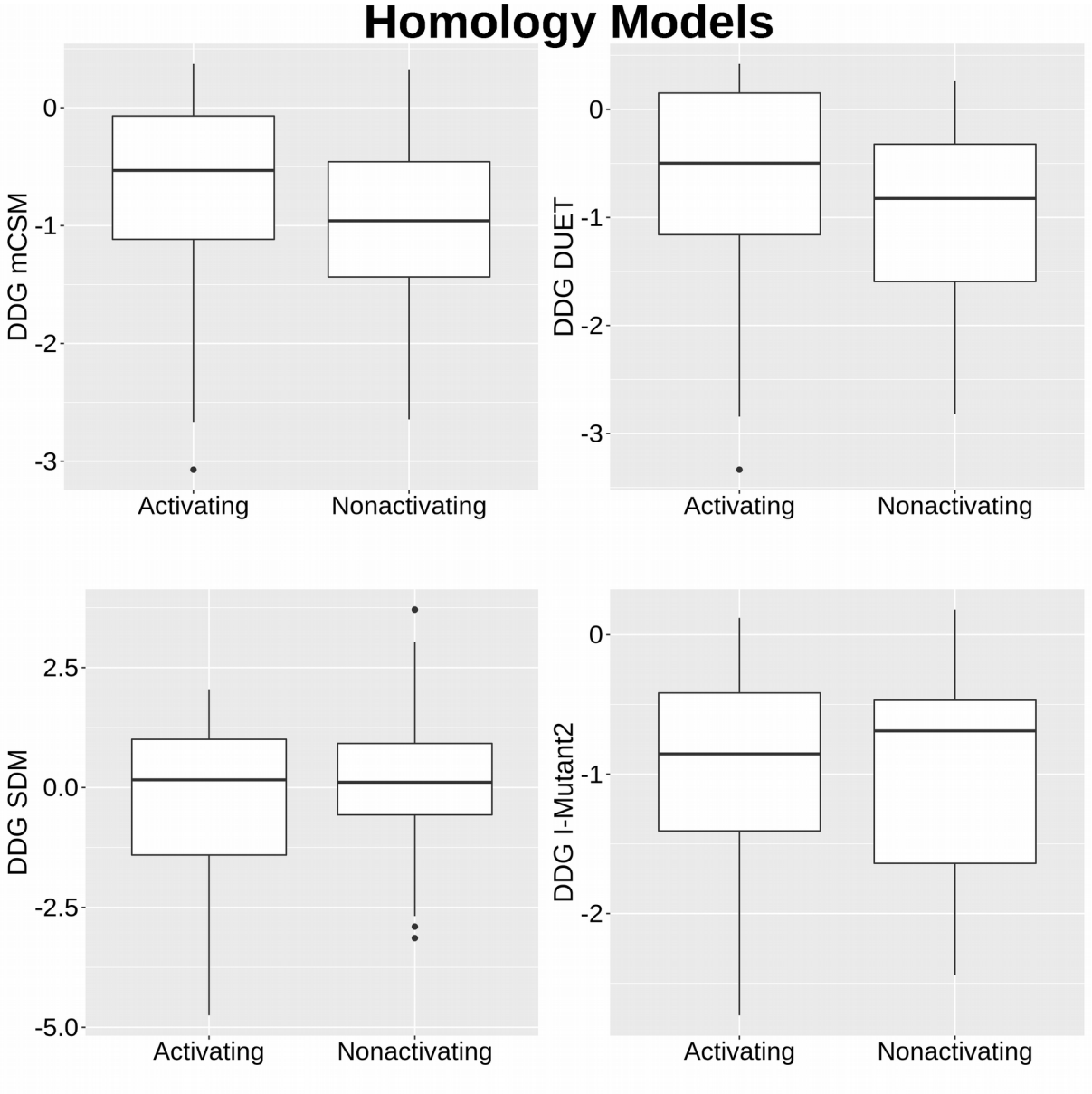


Figure 3 – Distribution of values for mCSM, SDM, DUET and I-Mutant2 for each validation status (Activating and Non-activating) on the set of mutations mapped on structures generated by homology models, also used to validate Kinact. T-test with 95% confidence interval was performance and no significant difference was observed for none of the scores.

In addition, as an attempt to compare Kinact predictive performance with these tools using AUC metric, we considered stability scores below 0 as non-activating and above 0 as activating. The performance of all methods on blind test set and also on a set with homology models are shown on the figure below.

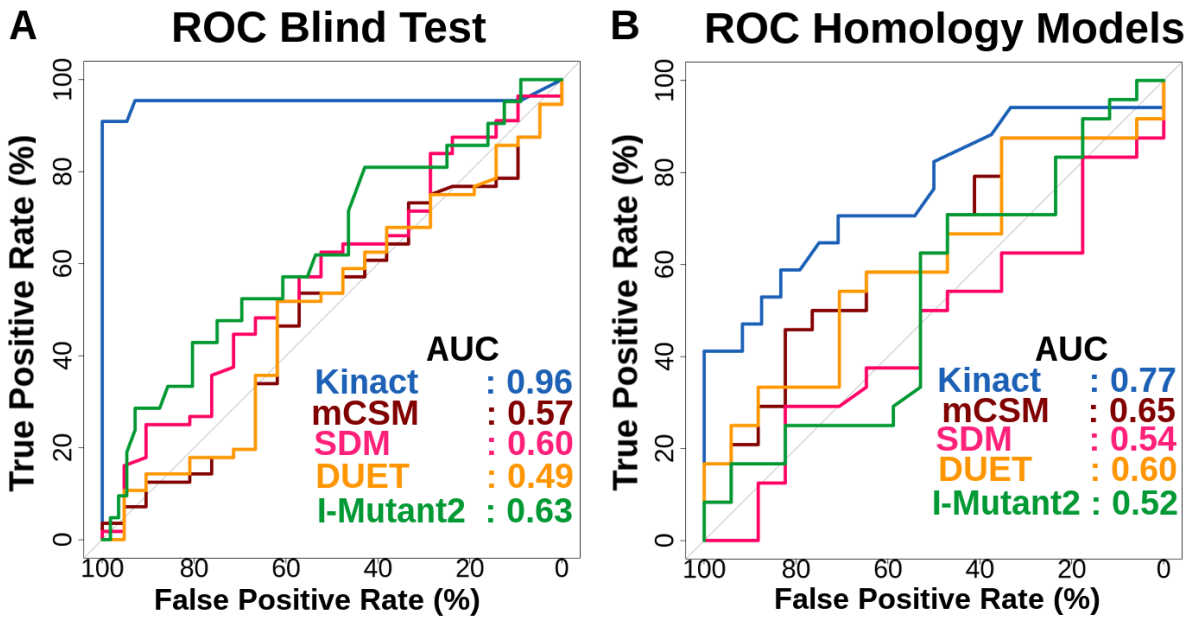


Figure 4 – Comparison of Kinact with well-established predictive methods that assess the effects of mutations on protein stability. Scores below 0 were considered non-activating and above 0 activating for mCSM, SDM, DUET and I-Mutant2.

Not surprisingly, Kinact outperformed all three methods on both test sets, given the fact that it was built specifically for the identification of gain of function mutations in protein kinases while the other tools were specifically built for the general purpose of understanding the effects of mutations on protein stability. In that sense, we strongly believe that Kinact and the other tools provide different information regarding the effects of mutations on proteins kinases and can be valuable tools for the study of these variants and the molecular mechanism of activation of these proteins. Moreover, given the importance of these variants in the context of many diseases, especially on the development of many types of cancer, we believe Kinact will be a useful tool to help identify and understand the role of these mutations.

TABLES

Table S1 - Description of categories of attributes generated. The table presents a short summary of the attributes and the data and tools used for their calculation.

Category of attributes	Attributes	Rely on	Tools
Wild-type residue environment	type of secondary structure, solvent accessibility, residue depth, dihedral angles, flexibility, minimum distance to catalytic sites and relative b-factor	Structure	Biopython, ENCoM, CSA
Wild-type residue interactions	clash, covalent, Van der Waals clash, vdw, proximal, hydrogen bond, weak hydrogen bond, halogen bond, ionic, metal complex, aromatic, hydrophobic, carbonyl, polar hydrogen bonds without angles, weka polar weak hydrogen bonds without angles	Structure	Arppegio
Structural signatures	distance patterns among the atoms of the structure based on graph modeling	Structure	mCSM
Stability change upon mutation	Variation of Gibbs Free Energy - ΔΔG	Structure	SDM, mCSM and DUET
Probability of damaging protein function	Tolerated or deleterious mutations that affects protein function	Sequence	Polyphen and SIFT
Pharmacophores	Pharmacophore differences based on protein sequence	Sequence	Pharmacophore difference,

Table S2 - Results for all classifiers trained with sequence-based features in each one of the mutation classes. Best performing model is highlighted.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,764	0,823	0,793	0,619	Activating
MLP	0,425	0,340	0,378	0,619	Non-activating
Regression (M5P)	0,730	0,915	0,812	0,617	Activating
Regression (M5P)	0,353	0,120	0,179	0,617	Non-activating
Random Forest	**0,775**	**0,877**	**0,823**	**0,769**	**Activating**
Random Forest	**0,659**	**0,482**	**0,557**	**0,769**	**Non-activating**
J48 Tree	0,750	0,877	0,809	0,531	Activating
J48 Tree	0,429	0,240	0,308	0,531	Non-activating

Table S3 - Results for all classifiers trained with structure-based features in each one of the mutation classes. Best performing model is highlighted.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,784	0,790	0,787	0.775	Activating
MLP	0,602	0,600	0,600	0,755	Non-activating
Regression (M5P)	0,810	0,843	0,826	0,793	Activating
Regression (M5P)	0,722	0,665	0,692	0,793	Non-activating
Random Forest	**0,873**	**0,862**	**0,867**	**0,833**	**Activating**
Random Forest	**0,715**	**0,680**	**0,697**	**0,833**	**Non-activating**
J48 Tree	0,794	0,812	0,802	0,713	Activating
J48 Tree	0,659	0,613	0,635	0,713	Non-activating

Table S4 - Results for all classifiers trained with the training test of mutations with structural and sequence-based features. Results are presented for both classes of mutations. Best performing model is highlighted.

Classifier	Precision	Recall	F-Measure	AUC	Class
MLP	0,884	0,931	0,907	0,885	Activating

MLP	0,791	0,680	0,731	0,885	Non-activating
Regression (M5P)	0,885	0,939	0,911	0,883	Activating
Regression (M5P)	0,810	0,680	0,739	0,833	Non-activating
Random Forest	**0,860**	**0,939**	**0,898**	**0,885**	**Activating**
Random Forest	**0,789**	**0,730**	**0,758**	**0,885**	**Non-activating**
J48 Tree	0,874	0,901	0,887	0,783	Activating
J48 Tree	0,717	0,660	0,688	0,783	Non-activating

FIGURES

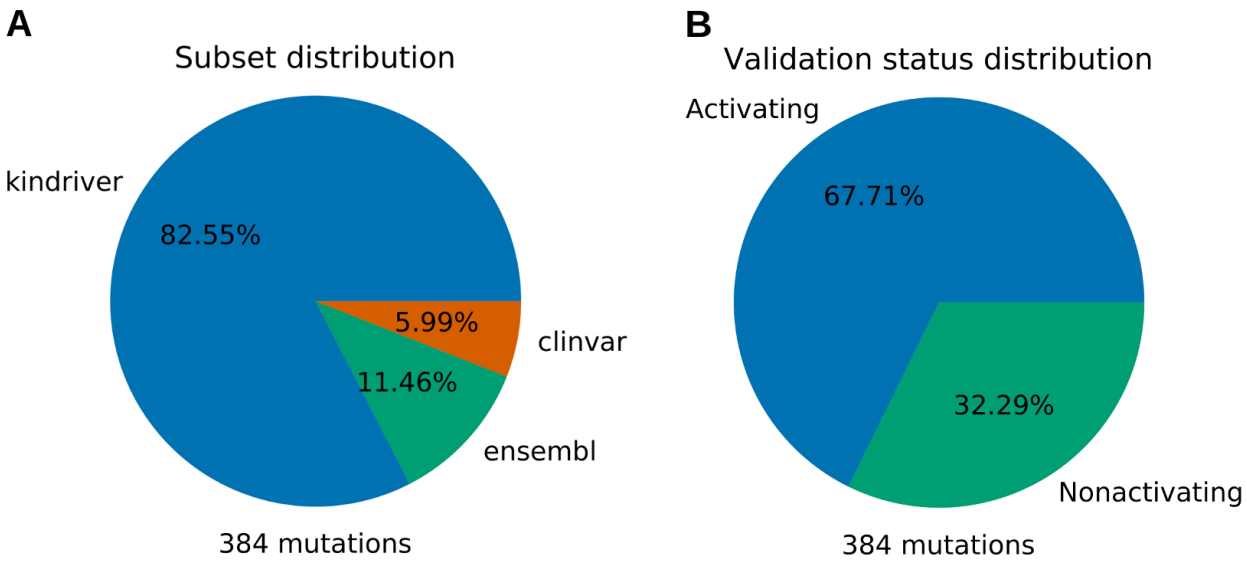


Figure S1 - Distribution of mutations and validations status. A) shows the distribution of mutations regarding the subsets of origin (Kin-driver, Clinvar and Ensembl) that were used for data collection. Most of the data was obtained from Kin-driver followed by Ensembl and Clinvar. B) depicts the distribution of validation status (activating, non-activating) of mutations across the entire data set.

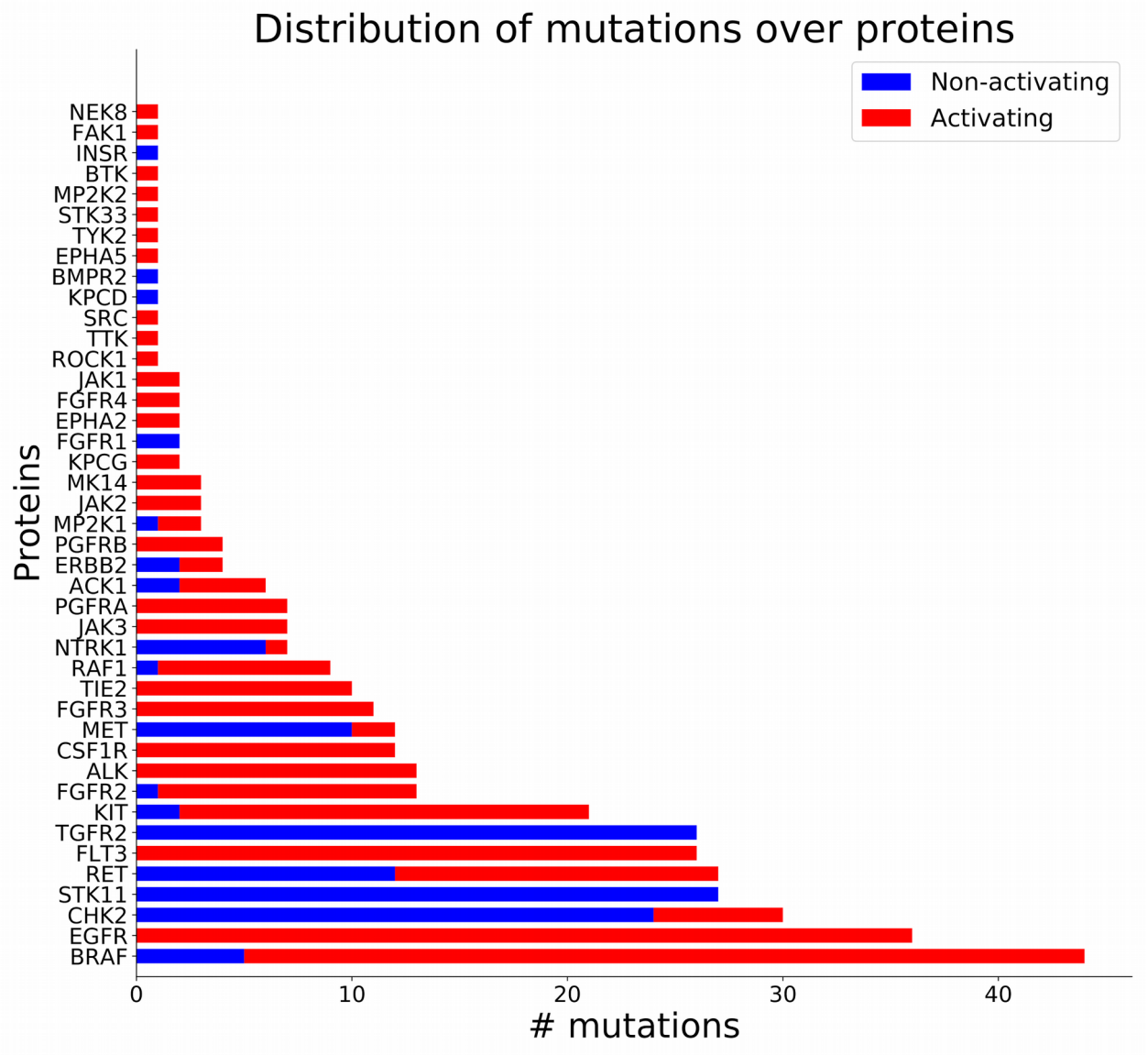


Figure S2 - Summary of number of mutations per protein within the data set of 384 mutations collected for this work. The top 3 proteins with highest number of mutations are BRAF, EGFR and CHK2 with 44, 36 and 30 mutations respectively. Bars are colored according to the class of mutations: blue bar indicates the amount of non-activating mutations and red bar indicates the amount of activating ones.

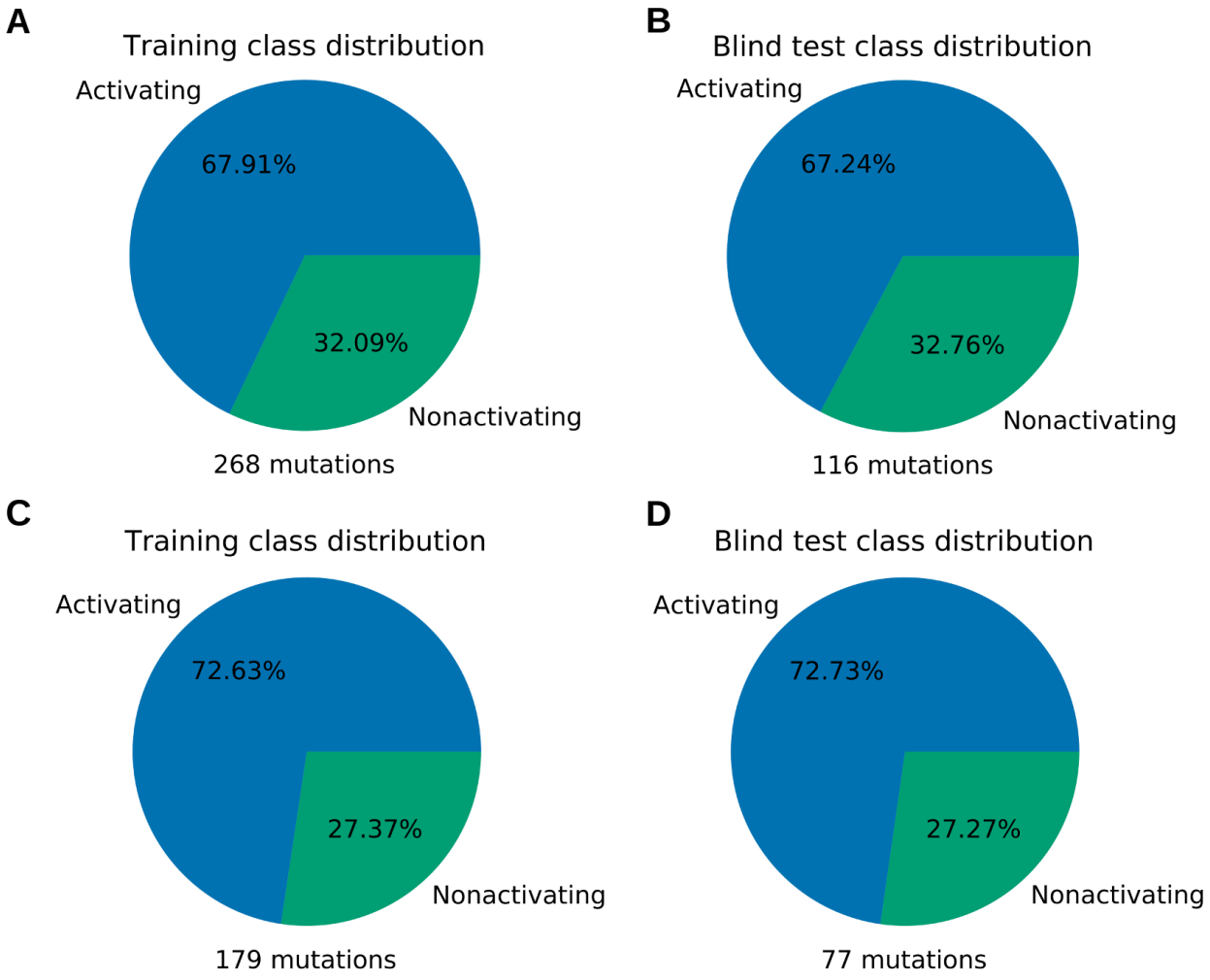


Figure S3 - Distribution of mutations in training and blind test sets used to build the final predictive model. The complete set of mutations was divided into two groups. The first comprising all 384 mutations identified during data collection. The second group contains only those mutations that had their region mapped into structures on the PDB. Each group is split into data for training and blind test of the machine learning algorithms. A) displays the distribution of the two classes of mutations (activating and non-activating) over the set of mutations used in training for data without structural mapping. B) presents the distribution of the two classes of mutations for the blind test on the same type of data. C) and D) introduces the class distribution for training and blind test, respectively, for mutations that had their region mapped into 3D structures of PDB.

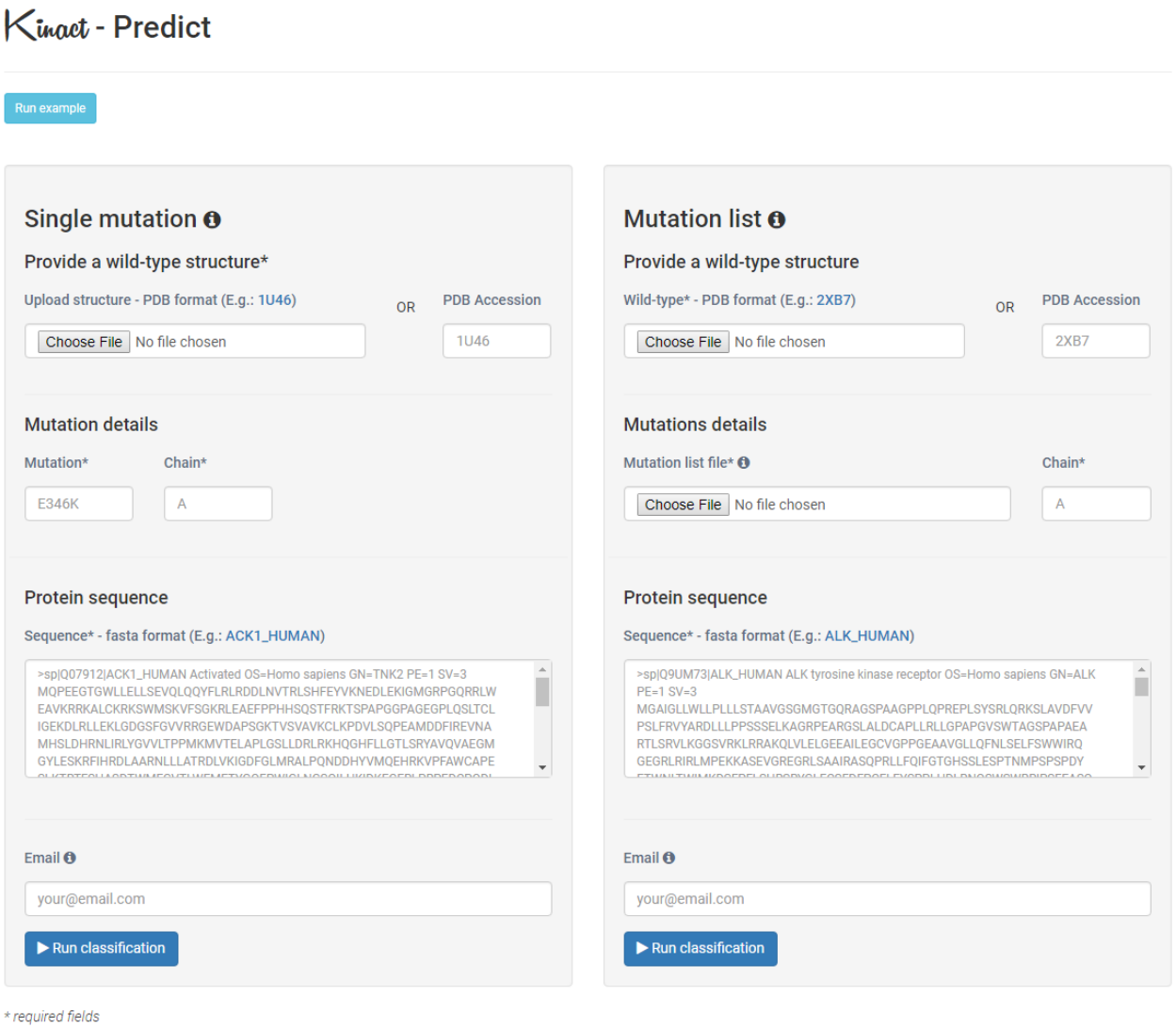


Figure S4 – Input page of Kinact. The server provides two different input options for the user. The "Single mutation" option allows users to predict whether a given mutation will lead to protein kinase activation or not. Users are required to provide a PDB file or PDB accession code of the kinase, the point mutation specified as a string containing the wild-type residue one-letter code, its corresponding residue number and the mutant residue one-letter code, and also the chain identifier of the wild-type residue. The primary sequence of the kinase of interest in fasta format is also required. The "Mutation list" option allows users to upload a list of mutations in a file for batch processing.

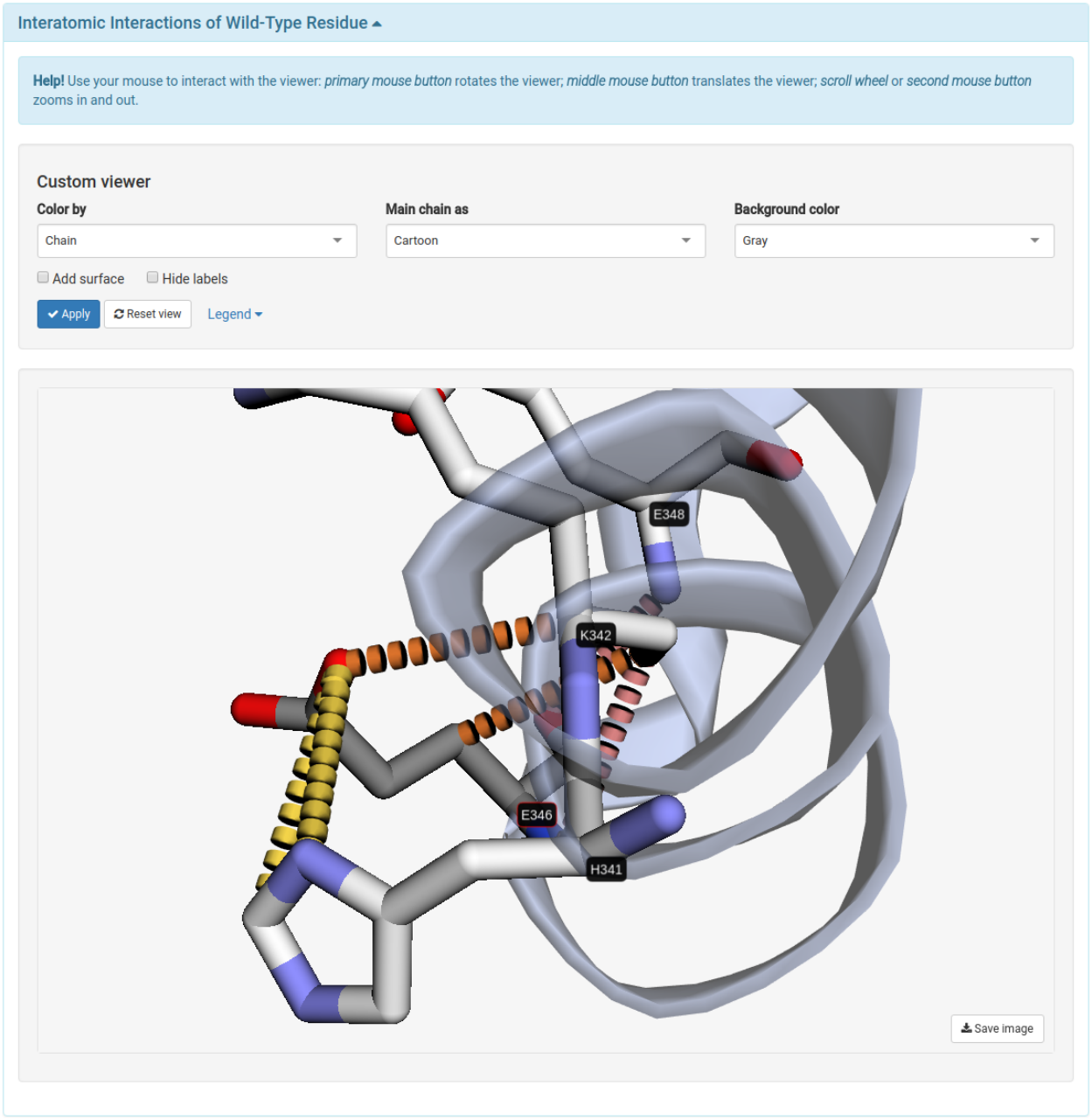


Figure S5 - Interatomic Interactions of Wild-Type Residue on Kinact Results page. By default, the molecule is displayed using the cartoon representation with the wild-type residue highlighted as stick and labeled, as well as the surrounding residues that make interactions with it according to Arpeggio. On the top of the page, a set of options allow the user to customise the viewer and a legend is also provided for the binding types.

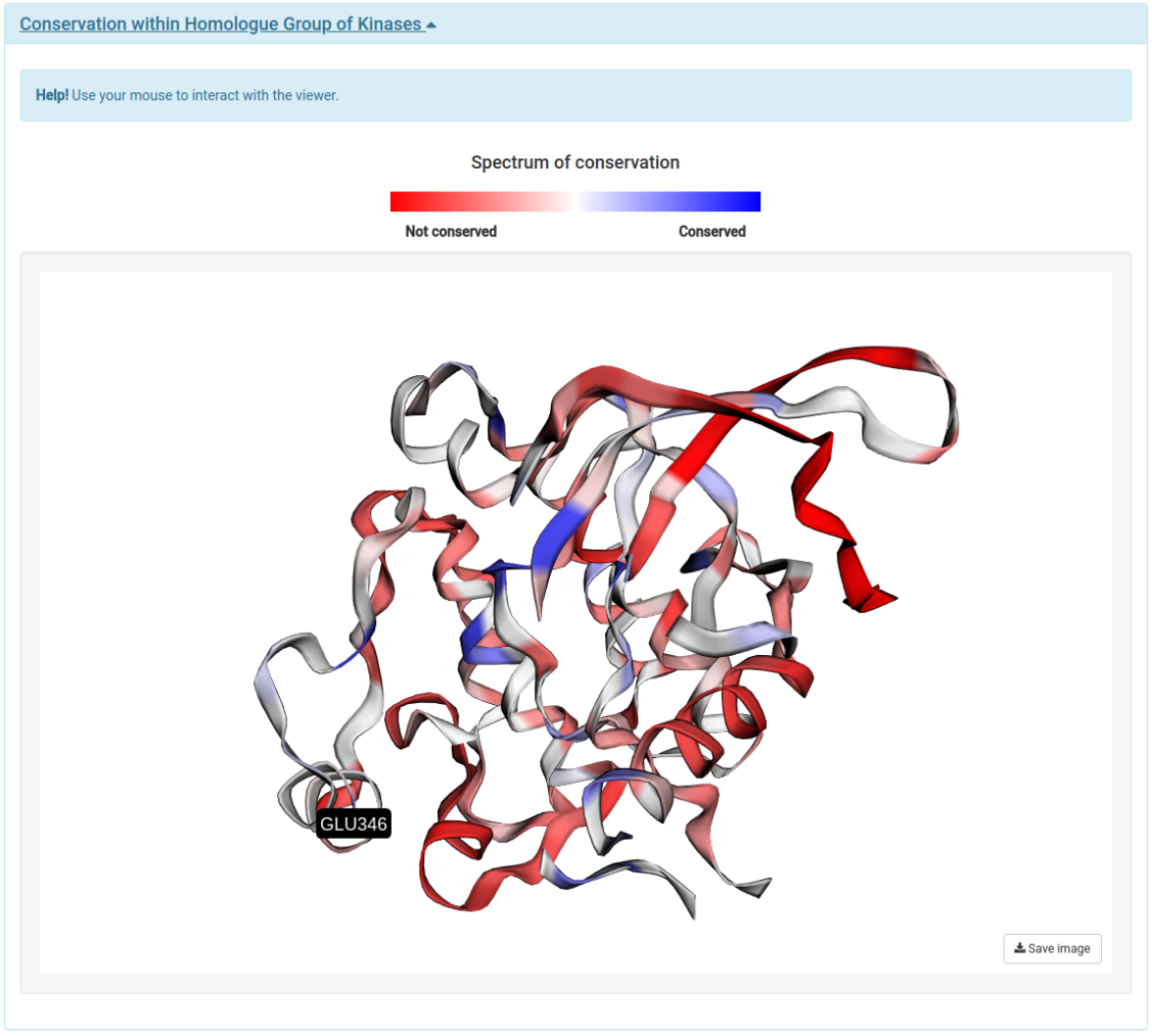


Figure S6 - Conservation analysis within Homologue Group of Kinases on Kinact Results page. 3D structure is colored according to residue conservation varying from red (not conserved) to blue (conserved). Wild-type residue is labeled.

Figure S7 - Multiple Sequence Alignment of Kinase Groupon Kinact Results page. This section shows the Multiple Sequence Alignment with the proteins of the group in which the submitted molecule was assigned according to Kinannote. All residues are colored by their type: Polar as pink, Hydrophobic as light green, Charged as blue and Sulphur as orange. Activating mutations are highlighted with red background. A legend is shown on top of the page.

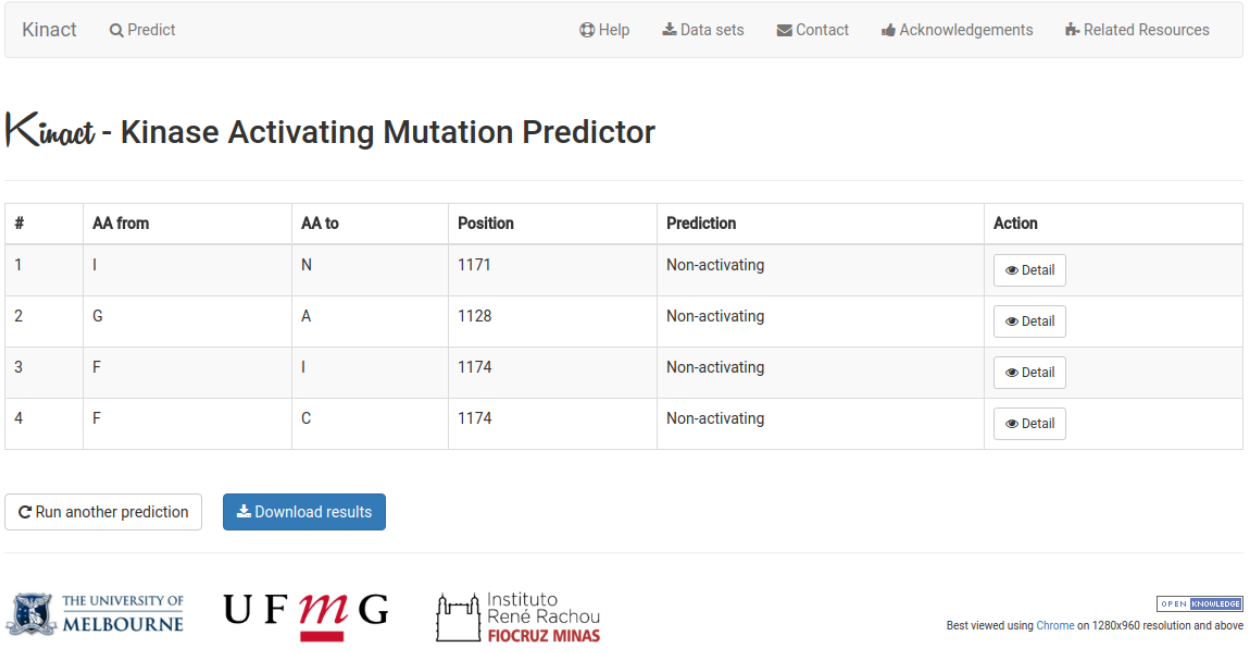


Figure S8 - Kinact Results page for a list of mutations. A table with the prediction outcome for each of the submitted mutations. The detail button allows users to perform analysis for each mutation individually, similar to those described for the 'Single Mutation' option.

REFERENCES

1. Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.
http://www.ncbi.nlm.nih.gov/pubmed/7063747
http://dx.doi.org/10.1148/radiology.143.1.7063747

2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10-18.
http://dx.doi.org/10.1145/1656274.1656278
http://www.ncbi.nlm.nih.gov/pmc/articles/1656278

3. Kotsiantis, S.B. (2007), *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, pp. 3-24.
4. Frank, E., Wang, Y., Inglis, S., Holmes, G. and Witten, I.H. (1998) Using Model Trees for Classification. *Machine Learning*, **32**, 63-76.
http://dx.doi.org/10.1023/a:1007421302149

5. Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.
http://dx.doi.org/10.1023/a:1010933404324

6. Powers, D.M.W. (2011) Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, **2**, 37-63.
http://dx.doi.org/citeulike-article-id:12882259

7. Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E. (2006) Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, **26**, 159-190.
http://dx.doi.org/10.1007/s10462-007-9052-3

8. Salzberg, S.L. (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, **16**, 235-240.
http://dx.doi.org/10.1007/bf00993309

9. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335-342.
http://www.ncbi.nlm.nih.gov/pubmed/24281696
http://dx.doi.org/10.1093/bioinformatics/btt691

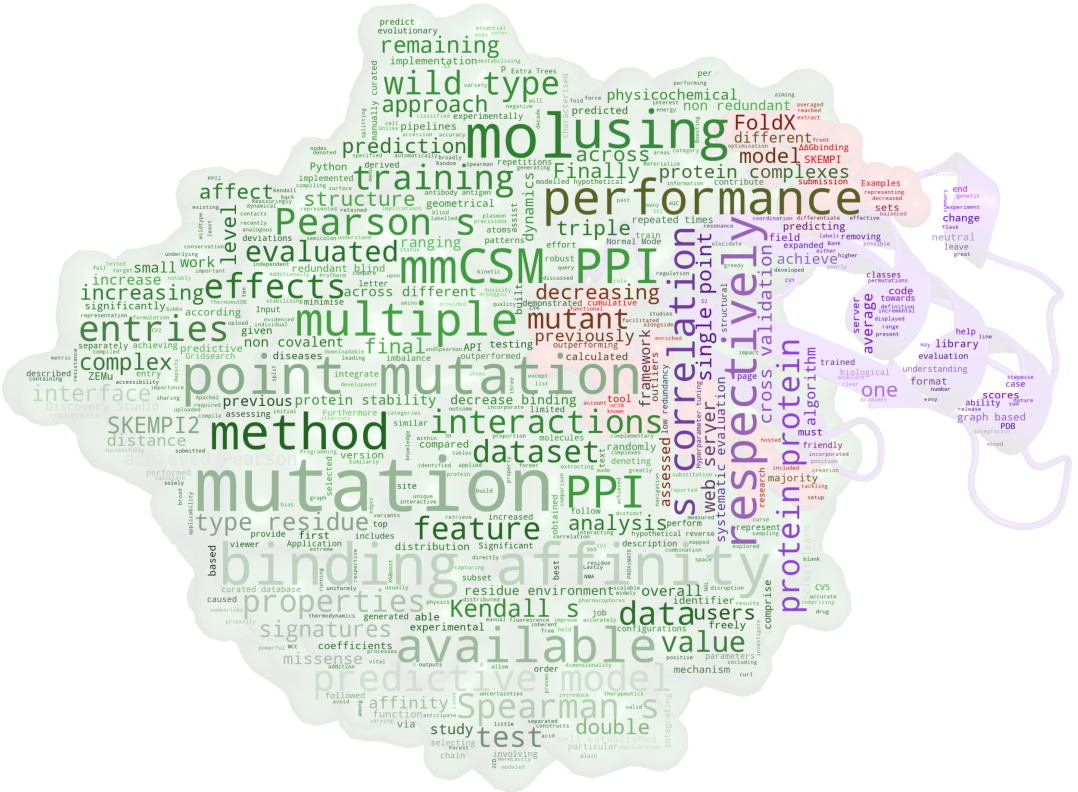
10. Pandurangan, A.P., Ochoa-Montano, B., Ascher, D.B. and Blundell, T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res*, **45**, W229-W235.
http://www.ncbi.nlm.nih.gov/pubmed/28525590
http://dx.doi.org/10.1093/nar/gkx439

11. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*, **42**, W314-319.
http://www.ncbi.nlm.nih.gov/pubmed/24829462
http://dx.doi.org/10.1093/nar/gku411

12. Capriotti, E., Fariselli, P. and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*, **33**, W306-310.
http://www.ncbi.nlm.nih.gov/pubmed/15980478
http://dx.doi.org/10.1093/nar/gki375

Chapter 7

Study of Effects of Multiple Mutations on Protein-protein Interactions



Nucleic Acids Research, 2021 ***1–8***
doi: 10.1093/nar/gkab273

mmCSM-PPI: predicting the effects of multiple point mutations on protein–protein interactions

Carlos H.M. Rodrigues[1,2,3], Douglas E.V. Pires[1,2,3,4,*] and David B. Ascher[1,2,3,5,*]

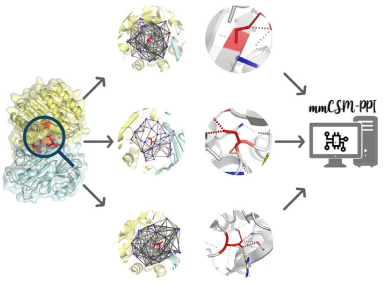
[1]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia, [2]Structural Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Victoria, Australia, [3]Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia, [4]School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia and [5]Department of Biochemistry, University of Cambridge, Cambridge, UK

Received January 27, 2021; Revised March 18, 2021; Editorial Decision April 2, 2021

ABSTRACT

Protein–protein interactions play a crucial role in all cellular functions and biological processes and mutations leading to their disruption are enriched in many diseases. While a number of computational methods to assess the effects of variants on protein–protein binding affinity have been proposed, they are in general limited to the analysis of single-point mutations and have been shown to perform poorly on independent test sets. Here, we present mmCSM-PPI, a scalable and effective machine learning model for accurately assessing changes in protein–protein binding affinity caused by single and multiple missense mutations. We expanded our well-established graph-based signatures in order to capture physicochemical and geometrical properties of multiple wild-type residue environments and integrate them with substitution scores and dynamics terms from normal mode analysis. mmCSM-PPI was able to achieve a Pearson's correlation of up to 0.75 (RMSE = 1.64 kcal/mol) under 10-fold cross-validation and 0.70 (RMSE = 2.06 kcal/mol) on a non-redundant blind test, outperforming existing methods. Our method is freely available as a user-friendly and easy-to-use web server and API at http://biosig.unimelb.edu.au/mmcsm_ppi.

GRAPHICAL ABSTRACT



INTRODUCTION

Protein-protein interactions (PPIs) are a vital mechanism for regulation and coordination of most biological processes within the cell (1,2). Missense mutations are known to directly contribute to function disruption and are enriched at their interacting interface in many diseases (3–7). The ability to elucidate the underlying mechanisms by which point mutations affect PPI interactions is therefore essential for understanding how to modulate these interactions and the development of therapeutics to target them.

Significant effort in the creation of manually curated databases compiling experimental data on the effects of mutations on protein stability and PPI binding affinity, most notably ThermomutDB (8), ProTherm (9), PROXiMATE (10) and SKEMPI (11,12), has greatly facilitated studies aiming to understand and predict how missense mutations affect PPIs. However, these have shown to perform poorly on independent test sets and are usually limited to predicting effects of single-point mutations. Furthermore, to the best of our knowledge, little effort has been made towards accessibility of these methods to help integration into other analysis pipelines.

We have shown previously that representing protein structure as a graph is a powerful method for extracting structural signatures as distance patterns (13). These com-

*To whom correspondence should be addressed. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au
Correspondence may also be addressed to Douglas E.V. Pires.Email: douglas.pires@unimelb.edu.au

pile geometrical and physicochemical properties which can further be mined and applied in a broad range of areas, such as predicting the effects of single-point missense mutations on protein stability (14–18), dynamics (16,17), interactions (15,19–25), genetic diseases (26–38) and drug resistance (39–53).

Here, we introduce mmCSM-PPI, a scalable and effective predictive model for assessing changes in PPI binding affinity caused by multiple missense mutations. We expanded our well-established graph-based signatures to allow for capturing physicochemical and geometrical properties of multiple wild-type residue environments, and integrate them with evolutionary scores, dynamics terms from Normal Mode Analysis (NMA) and non-covalent interactions for an accurate overall prediction (Figure 1).

MATERIALS AND METHODS

Data sets

The data used in this work was derived from SKEMPI2 (12), a manually curated database of experimental data on thermodynamics and kinetic parameters for wildtype and mutant protein–protein complexes which have been mapped to protein structures available on the Protein Data Bank (54). We were able to retrieve experimental information on 1721 multiple mutations, ranging from 2 to 27 point mutations, across 147 different protein–protein complexes (Supplementary Table S1). These had been primarily experimentally characterised by surface plasmon resonance and fluorescence methods (Supplementary Table S2 and Supplementary Figure S1).

Wild-type and mutant binding affinity parameters from SKEMPI2 were used to calculate the Gibbs free energy of binding as follows:

$$\Delta G^{\text{binding}} = RT\ln(K_{\text{D}})$$

where $R = 1.9872$ cal/K·mol is the ideal gas constant, T is the temperature (in K) and K_{D} is the affinity of the protein–protein complex.

The change in binding affinity upon mutation was calculated with the formulation previously described in SKEMPI2 and used in previous works:

$$\Delta\Delta G^{\text{binding}} = \Delta G^{\text{binding}}{}_{\text{WT}} - \Delta G^{\text{binding}}{}_{\text{MT}}$$

With positive values denoting mutations leading to an increased affinity and negative values denoting decreased binding affinity, given in kcal/mol. As shown in Supplementary Figure S2, the majority of entries in our dataset (1126) comprise double and triple mutants and for this work these were used as evidenced to train our predictive model. Furthermore, we explored the performance of our method on low-redundancy sets at complex and binding interface level according to the definition used in SKEMPI2. The remaining 595 multiple point mutations (2 neutral, 153 increasing and 440 decreasing affinity), ranging from 4–27 mutations, were held out and used as a non-redundant blind test at mutation level for performance comparison.

The distribution of $\Delta\Delta G^{\text{binding}}$ (Supplementary Figure S3A) depicts a clear bias towards mutations that decrease binding affinity ($\Delta\Delta G^{\text{binding}} < 0$ kcal/mol) in the training set. To minimize the imbalance nature of the dataset and how it would affect our predictive model, we also included modelled hypothetical reverse mutations in the training set (55,56). Unlike previous implementations, however, here we only modelled hypothetical reverse mutations for entries where -0.5 kcal/mol $< \Delta\Delta G^{\text{binding}} < 0.5$ kcal/mol to minimize uncertainties about the quality and biological implications of the modeled mutant structure (17). Therefore, the final training set used in this study includes 1344 entries, 12 neutral ($\Delta\Delta G^{\text{binding}} = 0$ kcal/mol), 347 increasing ($\Delta\Delta G^{\text{binding}} < 0$ kcal/mol) and 985 decreasing binding affinity ($\Delta\Delta G^{\text{binding}} > 0$ kcal/mol). All datasets used for training and test are freely available at http://biosig.unimelb.edu.au/mmcsm_ppi/data.

Graph-based signatures

Our graph-based structural signatures framework is a well-established approach used to represent physicochemical and geometrical properties of protein structure and small molecules. In the past decade, our method has been widely used for assessing the effects of single point-mutations on protein stability (14–16,18), PPI and antibody-antigen binding affinity (15,19,23,25), and small molecules toxicity (57–59). More recently, we have successfully expanded the applicability of our approach to investigate the impact of multiple point-mutations on protein stability (17) and on antibody-antigen binding affinity (24).

In this work, for each point-mutation, our signatures represent atoms of the wild-type residues as nodes and their interactions as edges, where their physicochemical properties are incorporated as labels according to amino acid residue properties (pharmacophores). The representation of the each wild-type residue environment is then used to extract distance patterns between atoms characterised by their properties and compiled in signatures as cumulative distributions. Finally, the cumulative distributions are averaged based on the number of point-mutations (Supplementary Figure S4).

Modelling multiple-mutation effects

Similarly to our previous implementation tackling the effects of single-point mutations on PPI binding affinity (15,23), here we also incorporate complementary features to account for the different mechanisms by which multiple point mutations may affect PPIs. However, in this study, we calculated the sum and average values of each property in order to model the effects of multiple mutations. All features generated can be broadly classified into 6 different categories: (i) dynamics, obtained via normal mode analysis (60), (ii) residue environment properties (61), (iii) conservation, obtained by using scores from substitution tables (62), (iv) non-covalent contacts involving wild-type residues (63), (v) wild-type inter-residue distance and (vi) predicted $\Delta\Delta G^{\text{binding}}$ for each single point mutation separately (23). A summary of features for each category is available in Supplementary Table S3.

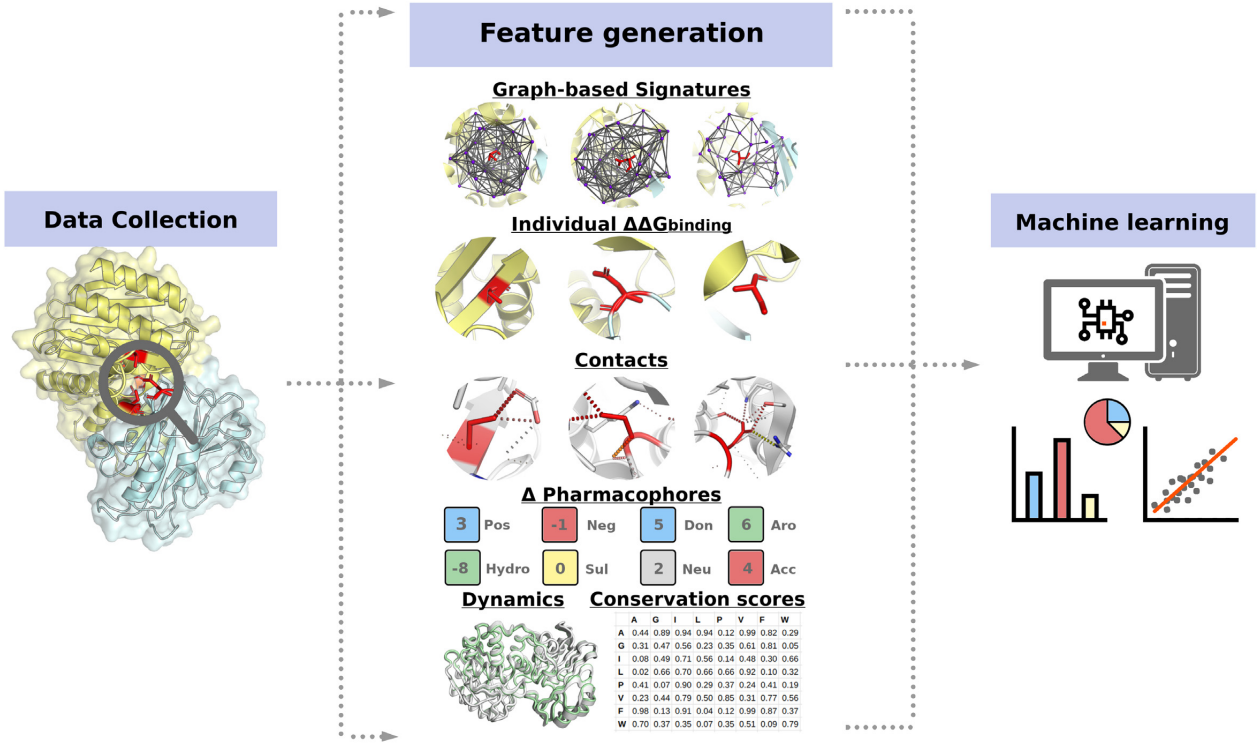


Figure 1. mmCSM-PPI methodology workflow. Experimental data on the effects of multiple missense mutations was collected from SKEMPI2 and mapped on their respective protein structures on the PDB. These were then used to generate physicochemical and geometrical properties in the form of graph-based signatures. In addition, six distinct types of complementary features were calculated to account for different mechanisms by which mutations may affect PPIs: (i) dynamic properties from NMA; (ii) wild-type residues environments; (iii) evolutionary and contact potential scores; (iv) non-covalent contacts; (v) wild-type inter-residue distances and (vi) the individual $\Delta\Delta G^{binding}$ for each point mutation. Feature selection was carried out with a stepwise greedy approach to avoid the curse of dimensionality and the best performing supervised learning algorithm was fine-tuned using the GridSearch function from the Scikit-learn Python library.

Machine learning

In this study we evaluated four distinct algorithms available on the scikit-learn Python library (64) on 10-fold cross-validation: Extra Trees, Random Forest, Gradient Boosting and XGBoost. The best performing algorithm used to build the final model was Extra Trees, based on different correlation coefficients (Pearson, Kendall and Spearman) and RMSE. Supplementary Table S4 summarises the performances of each algorithm. In order to avoid the curse of dimensionality and improve performance, we selected our features using an incremental stepwise greedy approach. Hyperparameter tuning was performed using the Gridsearch function also available on the scikit-learn library (Supplementary Table S5). Feature importance for the final predictive model is available on Supplementary Table S6. While two classes of features, graph-based signatures and individual mutation effects, were identified as contributing the most to the final model (as shown in Supplementary Table S7), their combination allowed for a significant increase in performance in the final model (P-value < 0.05), indicating they measure complementary aspects of mutation effects in PPIs.

WEB SERVER

We have implemented mmCSM-PPI as a user-friendly and freely available web server (http://biosig.unimelb.edu.au/mmcsm_ppi). The server front end was developed using Materialize framework version 1.0.0, and the back end was built using Python via the Flask framework (version 1.0.2). The web server is hosted on a Linux Server running Apache2.

Input

mmCSM-PPI can be used to either predict the effects of a list of mutations of interest or perform a systematic evaluation of all double and triple multiple mutations at a protein–protein interface (Supplementary Figure S5). In both cases, users are required to upload a file in PDB format or provide a valid PDB accession code with the structure of a protein–protein complex. For user-specified variants, mutations can be provided using a text field or uploaded as a plain text file with one multiple mutation per line. Each entry must be separated by a semicolon (;) and each point mutation must be represented as the chain identifier, blank space, the one-letter code for the wild-type, residue position and the one-letter code for the mutant. For the systematic evaluation option, users must provide a chain identifier from which interfaces will be automatically identified and all possible per-

mutations of double and triple mutations assessed. Examples and format descriptions are available in both submission page and help page via the top navigation menu.

An Application Programming Interface (API) to assist users in integrating our predictive tool into their research pipelines is also available. Input fields follow the same format previously described for our web server implementation. All jobs submitted are labelled with a unique identifier which is used to query the status of the job. A full description of the API, including examples using curl and Python are available at http://biosig.unimelb.edu.au/mmcsm_ppi/api.

Output

For both types of submissions, manual input and systematic evaluation, mmCSM-PPI outputs the predictions for all entries as a downloadable table where the predicted effects of multiple mutations on $\Delta\Delta G^{\text{binding}}$ is given in kcal/mol. For the systematic evaluation option, the server shows the top 100 increasing/decreasing affinity entries. Additionally, individual predictions for each point mutation is available, generated using mCSM-PPI2 (23), are also shown alongside the average distance among the wild-type residues. Finally, an interactive 3D viewer, built using the NGL viewer (65), allows for the analysis of non-covalent interactions involving wild-type residues for each point mutation, calculated using Arpeggio (63), for a particular entry. Users can alternate the residues and interactions being displayed by selecting different entries from the table (Supplementary Figure S6).

VALIDATION

Performance on cross-validation

We evaluated the performance of mmCSM-PPI across 5 different types of cross-validations on our training set. First, we randomly selected 80% of the data for training and remaining 20% for testing, repeated 100 times (CV1). Our method achieved Pearson's, Kendall's and Spearman's correlations of 0.87, 0.68 and 0.85 respectively, with small deviations across repetitions ($\sigma = 0.02$), and average RMSE of 1.41 kcal/mol ($\sigma = 0.21$). Using an analogous setup, but varying the proportion of data split for train and test (50% each set) (CV2), the performance was consistent with the previous experiment, and the predictive model achieved a Pearson's, Kendall's and Spearman's correlations of 0.86, 0.66 and 0.84 ($\sigma = 0.01$ for all coefficients), respectively (Figure 2A), and RMSE = 1.55 kcal/mol ($\sigma = 0.14$).

Since all the entries in our data set were not uniformly distributed across all protein–protein complexes (Supplementary Table S8), we evaluated the performance of our approach by randomly sampling up to 10 mutations per protein complex, repeated 10 times (generating 10 subsets), followed by randomly selecting 80% of entries for training and remaining 20% for testing, also repeated 10 times (CV3). For this type of cross-validation, our predictive model was able to achieve Pearson's, Kendall's and Spearman's correlations of 0.83, 0.63 and 0.81, again with small deviations over the repetitions ($\sigma = 0.03$) (Figure 2A), and average RMSE = 1.85 kcal/mol ($\sigma = 0.40$).

Finally, we assessed the robustness of mmCSM-PPI on low-redundancy sets at complex (CV4) and interface (CV5) levels. The former was implemented using leave-one-complex-out cross-validation, where all mutations for a particular complex were retained for test and the remaining for training the predictive model. Overall, our predictive model achieved Pearson's, Kendall's and Spearman's correlations of 0.76, 0.55 and 0.75 respectively, and RMSE of 1.59 kcal/mol (Figure 2B). On leave-one-binding-site-out (CV5), where all mutations for protein–protein complexes sharing similar binding sites, according to data on SKEMPI2, were used for testing and the remaining for training, our method was able to achieve Pearson's, Kendall's and Spearman's correlations of 0.73, 0.54 and 0.74, respectively (RMSE = 1.40 kcal/mol).

Blind test

While mmCSM-PPI was trained using a subset containing only double and triple mutants, the performance of our final model was further evaluated using a non-redundant blind set at the mutation level of experimentally measured effects of 595 constructs with at least four point mutations, also derived from SKEMPI2. Across this dataset, mmCSM-PPI achieved Pearson's, Kendall's and Spearman's correlation coefficients of 0.70, 0.48 and 0.64, respectively, and RMSE of 2.02 kcal/mol, significantly outperforming FoldX (66) and Discovery Studio (P-value < 0.05, Table 1). After removing 10% of outliers, the performance of our predictive model increased to 0.81, 0.55 and 0.73 for Pearson's, Kendall's and Spearman's correlations, respectively, and RMSE of 1.68 kcal/mol (Figure 2C). The majority of outliers (~70%) comprise mutations with extreme effects to PPI binding affinity (4 kcal/mol $< |\Delta\Delta G^{\text{binding}}| <$ 11 kcal/mol) and entries with 10 or more point mutations. Reassuringly, however, our final model demonstrated balanced predictive performance across both stabilising and destabilising mutations, achieving an overall accuracy of 87% and precisions of 74% and 89% on mutations that increase and decrease binding affinity, respectively.

Given the inherent imbalance between increasing and decreasing affinity mutations in the dataset, we further assessed the performance of our method on these respective classes separately. On mutations that decrease binding affinity, mmCSM-PPI achieves Pearson's, Kendall's and Spearman's correlations of 0.72, 0.46 and 0.64 respectively, with an RMSE = 1.67 kcal/mol, outperforming FoldX and Discovery Studio. For mutations that increase binding affinity all three methods show similar performance (Supplementary Table S9). Finally, we tested the ability to use the predicted $\Delta\Delta G^{\text{binding}}$ values from mmCSM-PPI to differentiate between mutations that increase from those that decrease binding affinity (Supplementary Table S10). Overall, our method has proven to be the most robust when compared with FoldX and Discovery Studio, achieving an AUC and MCC of 0.72 and 0.53, respectively, when evaluated on mutations where $|\Delta\Delta G^{\text{binding}}| < 1$ kcal/mol.

We further evaluated the generalisation capabilities of our model on another independent test set, non-redundant at the mutation level. Four hundred and ninety multiple-point mutations were randomly selected across 81 differ-

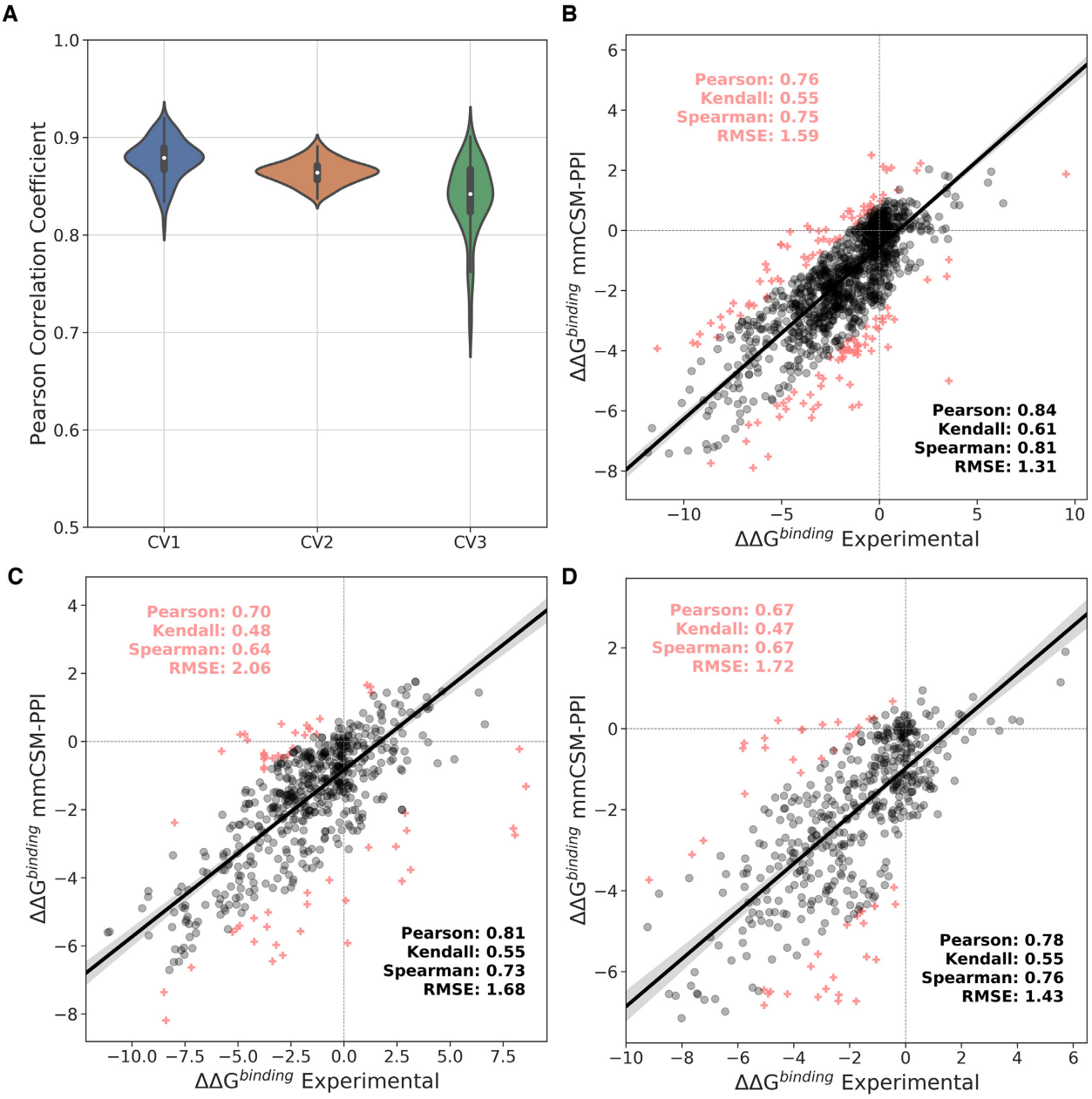


Figure 2. mmCSM-PPI performance on cross-validation and non-redundant blind-tests. (**A**) The performance of mmCSM-PPI on bootstrapped 5-fold cross validation (CV1), using 50% of the data as a blind test (CV2) and limiting the number of mutations per complex (CV3). The robustness of mmCSM-PPI was further assessed using low redundancy at the (**B**) complex level, (**C**) using all data with three or more mutations as a blind test, and (**D**) at the mutation level. Outliers are shown as red crosses.

Table 1. Performance comparison of mmCSM-PPI2 on a non-redundant blind test comprising entries with four or more mutations

Method	Pearson	Kendall	Spearman	RMSE (kcal/mol)	MCC	AUC
mmCSM-PPI	**0.70**	**0.48**	**0.64**	**2.02**	**0.53**	**0.72**
Discovery Studio	0.39*	0.29#	0.41+	3.07[a]	0.30	0.66
FoldX	0.39*	0.25#	0.37+	5.27[a]	0.22	0.61[b]

*P-value < 0.05 by Fisher r-to-z transformation test.
#$P < 0.05$ by transforming tau-to-r followed by Fisher r-to-z transformation.
+$P < 0.05$ by transforming rho-to-r followed by Fisher r-to-z transformation.
[a]$P < 0.05$ by Diebold–Mariano test.
[b]$P < 0.05$ by t-test.

ent PPI as a blind test, with the remaining being used for training purposes. Across the non-redundant blind test, mmCSM-PPI achieved Pearson's, Kendall's and Spearman's correlations of 0.67, 0.47 and 0.67, respectively (RMSE = 1.72 kcal/mol), performance consistent with previous independent tests, highlighting robustness of the method (Figure 2D).

The performance of mmCSM-PPI was compared to Discovery Studio and FoldX (Supplementary Table S11), which demonstrated that our approach significantly outperformed both in all metric evaluations (Supplementary Table S11). We also compared the performance of our method with ZEMu (67), a tool that uses a dynamical equilibration under a physics-based force field for a limited residue environment, followed by binding affinity evaluation with FoldX. In this case since ZEMu has only reported predictions for multiple mutations on the first version of SKEMPI, here we trained a predictive model with all double and triple mutants except for those available on the first release of SKEMPI. Therefore, the dataset used to compare the two methods comprises 272 entries (1 neutral, 52 increasing and 219 decreasing binding affinity) across 24 protein–protein complexes, ranging from 2 to 15 point mutations. mmCSM-PPI achieved Pearson's, Kendall's and Spearman's correlations of 0.73, 0.56 and 0.75 (RMSE = 1.72 kcal/mol), respectively, significantly higher (P-value < 0.05) than ZEMu (Pearson's, Kendall's and Spearman's correlations of 0.64, 0.46 and 0.65, respectively, and RMSE = 2.11 kcal/mol). On 90% of the dataset, our method achieves up to 0.83, 0.65 and 0.84 on Pearson's, Kendall's and Spearman's, respectively (RMSE = 1.49 kcal/mol).

CONCLUSION

Here, we present mmCSM-PPI, a web server that integrates our well-established graph-based signatures framework with evolutionary scores, dynamics properties and non-covalent interactions for accurately predicting changes in PPI binding affinity caused by multiple point mutations. Our method has shown to be robust when evaluated across different types of cross-validations and outperformed existing tools in a non-redundant blind test set. We anticipate mmCSM-PPI to be of great value for the study of how multiple mutations affect PPI binding affinity and to a variety of applications, ranging from protein functional analysis, optimisation of binding affinity and understanding the role of mutations in diseases. In addition, mmCSM-PPI includes an API to assist users when integrating our predictions into their research pipelines. Our method is freely available as a user-friendly and easy-to-use web server at http://biosig.unimelb.edu.au/mmcsm_ppi.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Melbourne Research Scholarship (to C.H.M.R.); Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council [MR/M026302/1 to D.B.A. and D.E.V.P.]; Jack Brockhoff Foundation [JBF 4186, 2016]; Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405]; Victorian Government's Operational Infrastructure Support Program (in part). Funding for open access charge: MRC.
Conflict of interest statement. None declared.

REFERENCES

1. Stumpf,M.P., Thorne,T., de Silva,E., Stewart,R., An,H.J., Lappe,M. and Wiuf,C. (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 6959–6964.
2. Blaszczyk,M., Harmer,N.J., Chirgadze,D.Y., Ascher,D.B. and Blundell,T.L. (2015) Achieving high signal-to-noise in cell regulatory systems: Spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. *Prog. Biophys. Mol. Biol.*, **118**, 103–111.
3. David,A., Razali,R., Wass,M.N. and Sternberg,M.J. (2012) Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.*, **33**, 359–363.
4. Engin,H.B., Kreisberg,J.F. and Carter,H. (2016) Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PLoS One*, **11**, e0152929.
5. Jubb,H., Blundell,T.L. and Ascher,D.B. (2015) Flexibility and small pockets at protein–protein interfaces: new insights into druggability. *Prog. Biophys. Mol. Biol.*, **119**, 2–9.
6. Jubb,H.C., Pandurangan,A.P., Turner,M.A., Ochoa-Montano,B., Blundell,T.L. and Ascher,D.B. (2017) Mutations at protein–protein interfaces: Small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Biol.*, **128**, 3–13.
7. Ascher,D.B., Jubb,H.C., Pires,D.E.V., Ochi,T., Higueruelo,A. and Blundell,T.L. (2015) In Scapin,G., Patel,D. and Arnold,E. (eds.), *Multifaceted Roles of Crystallography in Modern Drug Discovery*. Springer Netherlands, pp. 141–163.
8. Xavier,J.S., Nguyen,T.B., Karmarkar,M., Portelli,S., Rezende,P.M., Velloso,J.P.L., Ascher,D.B. and Pires,D.E.V. (2021) ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Res.*, **49**, D475–D479.
9. Kumar,M.D., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–206.
10. Jemimah,S., Yugandhar,K. and Michael Gromiha,M. (2017) PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, **33**, 2787–2788.
11. Moal,I.H. and Fernandez-Recio,J. (2012) SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.
12. Jankauskaite,J., Jimenez-Garcia,B., Dapkunas,J., Fernandez-Recio,J. and Moal,I.H. (2019) SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, **35**, 462–469.
13. Pires,D.E., de Melo-Minardi,R.C., dos Santos,M.A., da Silveira,C.H., Santoro,M.M. and Meira,W. Jr (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, **12**, S12.
14. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–319.
15. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
16. Rodrigues,C.H., Pires,D.E. and Ascher,D.B. (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.*, **46**, W350–W355.
17. Rodrigues,C.H.M., Pires,D.E.V. and Ascher,D.B. (2021) DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.*, **30**, 60–69.
18. Pires,D.E.V., Rodrigues,C.H.M. and Ascher,D.B. (2020) mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res.*, **48**, W147–W153.

19. Pires,D.E. and Ascher,D.B. (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.*, **44**, W469–473.
20. Pires,D.E. and Ascher,D.B. (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res.*, **44**, W557–561.
21. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.*, **6**, 29575.
22. Pires,D.E.V. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.
23. Rodrigues,C.H.M., Myung,Y., Pires,D.E.V. and Ascher,D.B. (2019) mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.*, **47**, W338–W344.
24. Myung,Y., Pires,D.E.V. and Ascher,D.B. (2020) mmCSM-AB: guiding rational antibody engineering through multiple point mutations. *Nucleic Acids Res.*, **48**, W125–W131.
25. Myung,Y., Rodrigues,C.H.M., Ascher,D.B. and Pires,D.E.V. (2020) mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics*, **36**, 1453–1459.
26. Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.
27. Usher,J.L., Ascher,D.B., Pires,D.E., Milan,A.M., Blundell,T.L. and Ranganath,L.R. (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. *JIMD Rep.*, **24**, 3–11.
28. Nemethova,M., Radvanszky,J., Kadasi,L., Ascher,D.B., Pires,D.E., Blundell,T.L., Porfirio,B., Mannoni,A., Santucci,A., Milucci,L. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur. J. Hum. Genet.*, **24**, 66–72.
29. Pires,D.E., Chen,J., Blundell,T.L. and Ascher,D.B. (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.*, **6**, 19848.
30. Casey,R.T., Ascher,D.B., Rattenberry,E., Izatt,L., Andrews,K.A., Simpson,H.L., Challis,B., Park,S.M., Bulusu,V.R., Lalloo,F. *et al.* (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol Genet. Genomic Med.*, **5**, 237–250.
31. Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B., Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom. Med.*, **2**, 7.
32. Hnizda,A., Fabry,M., Moriyama,T., Pachl,P., Kugler,M., Brinsa,V., Ascher,D.B., Carroll,W.L., Novak,P., Zaliova,M. *et al.* (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. *Leukemia*, **32**, 1393–1403.
33. Rodrigues,C.H., Ascher,D.B. and Pires,D.E. (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. *Nucleic Acids Res.*, **46**, W127–W132.
34. Ascher,D.B., Spiga,O., Sekelska,M., Pires,D.E.V., Bernini,A., Tiezzi,M., Kralovicova,J., Borovska,I., Soltysova,A., Olsson,B. *et al.* (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.*, **27**, 888–902.
35. Bayley,J.P., Bausch,B., Rijken,J.A., van Hulsteijn,L.T., Jansen,J.C., Ascher,D., Pires,D.E.V., Hes,F.J., Hensen,E.F., Corssmit,E.P.M. *et al.* (2020) Variant type is associated with disease characteristics in SDHB, SDHC and SDHD-linked phaeochromocytoma-paraganglioma. *J. Med. Genet.*, **57**, 96–103.
36. Hildebrand,J.M., Kauppi,M., Majewski,I.J., Liu,Z., Cox,A.J., Miyake,S., Petrie,E.J., Silk,M.A., Li,Z., Tanzer,M.C. *et al.* (2020) A missense mutation in the MLKL brace region promotes lethal neonatal inflammation and hematopoietic dysfunction. *Nat. Commun.*, **11**, 3150.
37. Jatana,N., Ascher,D.B., Pires,D.E.V., Gokhale,R.S. and Thukral,L. (2020) Human LC3 and GABARAP subfamily members achieve functional specificity via specific structural modulations. *Autophagy*, **16**, 239–255.
38. Trezza,A., Bernini,A., Langella,A., Ascher,D.B., Pires,D.E.V., Sodi,A., Passerini,I., Pelo,E., Rizzo,S., Niccolai,N. *et al.* (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest. Ophthalmol. Vis. Sci.*, **58**, 5320–5328.
39. Ascher,D.B., Wielens,J., Nero,T.L., Doughty,L., Morton,C.J. and Parker,M.W. (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci. Rep.*, **4**, 4765.
40. Hawkey,J., Ascher,D.B., Judd,L.M., Wick,R.R., Kostoulias,X., Cleland,H., Spelman,D.W., Padiglione,A., Peleg,A.Y. and Holt,K.E. (2018) Evolution of carbapenem resistance in *Acinetobacter baumannii* during a prolonged infection. *Microbial Genomics*, **4**, http://dx.doi.org/doi:10.1099/mgen.0.000165.
41. Holt,K.E., McAdam,P., Thai,P.V.K., Thuong,N.T.T., Ha,D.T.M., Lan,N.N., Lan,N.H., Nhu,N.T.Q., Hai,H.T., Ha,V.T.N. *et al.* (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.*, **50**, 849–856.
42. Karmakar,M., Globan,M., Fyfe,J.A.M., Stinear,T.P., Johnson,P.D.R., Holmes,N.E., Denholm,J.T. and Ascher,D.B. (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.*, **198**, 541–544.
43. Portelli,S., Phelan,J.E., Ascher,D.B., Clark,T.G. and Furnham,N. (2018) Understanding molecular consequences of putative drug resistant mutations in *Mycobacterium tuberculosis*. *Sci. Rep.*, **8**, 15356.
44. Vedithi,S.C., Malhotra,S., Das,M., Daniel,S., Kishore,N., George,A., Arumugam,S., Rajan,L., Ebenezer,M., Ascher,D.B. *et al.* (2018) Structural implications of Mutations Conferring Rifampin Resistance in Mycobacterium leprae. *Sci. Rep.*, **8**, 5016.
45. Karmakar,M., Rodrigues,C.H.M., Holt,K.E., Dunstan,S.J., Denholm,J. and Ascher,D.B. (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS One*, **14**, e0217169.
46. Karmakar,M., Rodrigues,C.H.M., Horan,K., Denholm,J.T. and Ascher,D.B. (2020) Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep.*, **10**, 1875.
47. Pires,D.E.V., Stubbs,K.A., Mylne,J.S. and Ascher,D.B. (2020) Designing safe and potent herbicides with the cropCSM online resource. bioRxiv, 2020.2011.2001.364240. http://dx.doi.org/10.1101/2020.11.01.364240, 02 November 2020, preprint: not peer reviewed.
48. Portelli,S., Myung,Y., Furnham,N., Vedithi,S.C., Pires,D.E.V. and Ascher,D.B. (2020) Prediction of rifampicin resistance beyond the RRDR using structure-based machine learning approaches. *Sci. Rep.*, **10**, 18120.
49. Vedithi,S.C., Rodrigues,C.H.M., Portelli,S., Skwark,M.J., Das,M., Ascher,D.B., Blundell,T.L. and Malhotra,S. (2020) Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of RNA polymerase in *Mycobacterium leprae*. *Comput Struct Biotechnol J*, **18**, 271–286.
50. Portelli,S., Olshansky,M., Rodrigues,C.H.M., D'Souza,E.N., Myung,Y., Silk,M., Alavi,A., Pires,D.E.V. and Ascher,D.B. (2020) Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. *Nat. Genet.*, **52**, 999–1001.
51. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res.*, **43**, D387–391.
52. Tunstall,T., Portelli,S., Phelan,J., Clark,T.G., Ascher,D.B. and Furnham,N. (2020) Combining structure and genomics to understand antimicrobial resistance. *Comput Struct Biotechnol J*, **18**, 3377–3394.
53. Vedithi,S.C., Malhotra,S., Skwark,M.J., Munir,A., Acebron-Garcia-De-Eulate,M., Waman,V.P., Alsulami,A., Ascher,D.B. and Blundell,T.L. (2020) HARP: a database of structural impacts of systematic missense mutations in drug targets of Mycobacterium leprae. *Comput. Struct. Biotechnol. J.*, **18**, 3692–3704.
54. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
55. Pandurangan,A.P., Ochoa-Montano,B., Ascher,D.B. and Blundell,T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.*, **45**, W229–W235.

56. Thiltgen,G. and Goldstein,R.A. (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One*, **7**, e46084.
57. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem.*, **58**, 4066–4072.
58. Kaminskas,L.M., Pires,D.E.V. and Ascher,D.B. (2019) dendPoint: a web resource for dendrimer pharmacokinetics investigation and prediction. *Sci. Rep.*, **9**, 15465.
59. Pires,D.E.V. and Ascher,D.B. (2020) mycoCSM: using graph-based signatures to identify safe potent hits against mycobacteria. *J. Chem. Inf. Model.*, **60**, 3450–3456.
60. Grant,B.J., Rodrigues,A.P., ElSawy,K.M., McCammon,J.A. and Caves,L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.
61. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
62. Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.
63. Jubb,H.C., Higueruelo,A.P., Ochoa-Montano,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.
64. Pedregosa,F., Varoquaux,G.l., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-Learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
65. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
66. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
67. Dourado,D.F. and Flores,S.C. (2014) A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins*, **82**, 2681–2690.

SUPPLEMENTARY MATERIAL

mmCSM-PPI: predicting the effects of multiple point mutations on protein-protein interactions

Carlos H. M. Rodrigues [1,2,3], Douglas E. V. Pires [1,2,3,4,*], David B. Ascher [1,2,3,5,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria
[2] Structural Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Victoria
[3] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria
[4] School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria
[5] Department of Biochemistry, University of Cambridge, Cambridge, UK

[*]To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

Extracting Binding Free Energy values from FoldX and Discovery Studio

In order to extract the effects of mutations on binding free energy from FoldX, we first generated mutant structures using the command BuildModel and then generated ΔGBinding for wild-type and mutant structures using the function AnalyseComplex. Information on the groups of molecules (chains) participating in the interaction and the change in binding free energy was then calculated using the definitions available in SKEMPI2:

$$\Delta\Delta G^{binding} = \Delta G^{binding}_{WT} - \Delta G^{binding}_{MT}$$

For Discovery Studio, we obtained values for changes in binding free energy using the Pipeline Plot 2018 module, running on CHARMM force-field and default configurations.

TABLES

Table S1 - Description of PPI structures present in the dataset used to build mmCSM-PPI. The majority of structures have been generated using X-ray crystallography and a minor fraction using Nuclear Magnetic Resonance (NMR). For the latter, resolution values have been set as “NA”.

PDB ID	Description	Experiment	Resolution
1cz8	Vascular endothelial growth factor in complex with an affinity matured antibody	x-ray diffraction	2.40
5m2o	R. flavefaciens' third scab cohesin in complex with a group 1 dockerin	x-ray diffraction	1.26
4l3e	The complex between high affinity tcr dmf5(alpha-d26y,beta-l98w) and human class i mhc hla-a2 with the bound mart-1(26-35)(a27l) peptide	x-ray diffraction	2.56
1mhp	Crystal structure of a chimeric alpha1 integrin i-domain in complex with the fab fragment of a humanized neutralizing antibody	x-ray diffraction	2.80
2c5d	Structure of a minimal gas6-axl complex	x-ray diffraction	3.30
1yqv	The crystal structure of the antibody fab hyhel5 complex with lysozyme at 1.7a resolution	x-ray diffraction	1.70

3idx	Crystal structure of hiv-gp120 core in complex with cd4-binding site antibody b13, space group c222	x-ray diffraction	2.50
2oob	Crystal structure of the uba domain from cbl-b ubiquitin ligase in complex with ubiquitin	x-ray diffraction	1.90
4gnk	Crystal structure of galphaq in complex with full-length human plcbeta3	x-ray diffraction	4.00
1qab	The structure of human retinol binding protein with its carrier protein transthyretin reveals interaction with the carboxy terminus of rbp	x-ray diffraction	3.20
1c1y	Crystal structure of rap.gmppnp in complex with the ras- binding-domain of c-raf1 kinase (rafrbd).	x-ray diffraction	1.90
4b0m	Complex of the caf1an usher domain, caf1m chaperone and caf1 subunit from yersinia pestis	x-ray diffraction	1.80
4g2v	Structure complex of lgn binding with frmpd1	x-ray diffraction	2.40
1tm1	Crystal structure of the complex of subtilisin bpn' with chymotrypsin inhibitor 2	x-ray diffraction	1.70
3hg1	Germline-governed recognition of a cancer epitope by an immunodominant human t cell receptor	x-ray diffraction	3.00
1dqj	Crystal structure of the anti-lysozyme antibody hyhel-63 complexed with hen egg white lysozyme	x-ray diffraction	2.00
1n8z	Crystal structure of extracellular domain of human her2 complexed with herceptin fab	x-ray diffraction	2.52
1ycs	P53-53bp2 complex	x-ray diffraction	2.20
1y4a	Crystal structure of the complex of subtilisin bpn' with chymotrypsin inhibitor 2 m59r/e60s mutant	x-ray diffraction	1.60
1dan	Complex of active site inhibited human blood coagulation factor viia with human recombinant soluble tissue factor	x-ray diffraction	2.00
1yy9	Structure of the extracellular domain of the epidermal growth factor receptor in complex with the fab fragment of cetuximab/erbitux/imc- c225	x-ray diffraction	2.61
4jfd	Preservation of peptide specificity during tcr-mhc contact dominated affinity enhancement of a melanoma-specific tcr	x-ray diffraction	2.46
4jfe	Preservation of peptide specificity during tcr-mhc contact dominated affinity enhancement of a melanoma-specific tcr	x-ray diffraction	2.70
1lfd	Crystal structure of the active ras protein complexed with the ras- interacting domain of ralgds	x-ray diffraction	2.10
4ra0	An engineered axl 'decoy receptor' effectively silences the gas6-axl signaling axis	x-ray diffraction	3.07
3uih	Crystal structure of human survivin in complex with smac/diablo(1-15) peptide	x-ray diffraction	2.40
2pye	Crystal structures of high affinity human t-cell receptors bound to pmhc revealnative diagonal binding geometry tcr clone c5c1 complexed with mhc	x-ray diffraction	2.30

3bdy	Dual specific bh1 fab in complex with vegf	x-ray diffraction	2.60
1ak4	Human cyclophilin a bound to the amino-terminal domain of hiv-1 capsid	x-ray diffraction	2.36
1gc1	Hiv-1 gp120 core complexed with cd4 and a neutralizing human antibody	x-ray diffraction	2.50
3be1	Dual specific bh1 fab in complex with the extracellular domain of her2/erbb-2	x-ray diffraction	2.90
4mnq	Tcr-peptide specificity overrides affinity enhancing tcr-mhc interactions	x-ray diffraction	2.74
1E50	Aml1/cbfbeta complex	x-ray diffraction	2.60
2dvw	Structure of the oncoprotein gankyrin in complex with s6 atpase of the 26s proteasome	x-ray diffraction	2.30
3qib	Crystal structure of the 2b4 tcr in complex with mcc/i-ek	x-ray diffraction	2.70
1a4y	Ribonuclease inhibitor-angiogenin complex	x-ray diffraction	2.00
1mq8	Crystal structure of alphal i domain in complex with icam-1	x-ray diffraction	3.30
1gua	Human rap1a, residues 1-167, double mutant (e30d,k31e) complexed with gppnhp and the ras-binding-domain of human c-raf1, residues 51-131	x-ray diffraction	2.00
1rew	Structural refinement of the complex of bone morphogenetic protein 2 and its type ia receptor	x-ray diffraction	1.86
4jeu	Crystal structure of munc18a and syntaxin1 with native n-terminus complex	x-ray diffraction	3.20
1efn	Hiv-1 nef protein in complex with r96i mutant fyn sh3 domain	x-ray diffraction	2.50
2o3b	Crystal structure complex of nuclease a (nuca) with intra-cellular inhibitor nuia	x-ray diffraction	2.30
1kne	Chromo domain of hp1 complexed with histone h3 tail containing trimethyllysine 9	x-ray diffraction	2.40
2b2x	Vla1 rdeltah i-domain complexed with a quadruple mutant of the aqc2 fab	x-ray diffraction	2.20
4uyq	High resolution structure of the third cohesin scac in complex with the scab dockerin with a mutation in the c-terminal helix (in to si) from acetivibrio cellulolyticus displaying a type i interaction.	x-ray diffraction	1.81
4ftv	The complex between the high affinity version of a6 tcr (a6c134) and human class i mhc hla-a2 with the bound tax nonameric peptide	x-ray diffraction	2.74
1mah	Fasciculin2-mouse acetylcholinesterase complex	x-ray diffraction	3.20
3d3v	The complex between tcr a6 and human class i mhc hla-a2 with the modified htlv-1 tax (y5(3,4-difluorophenylalanine)) peptide	x-ray diffraction	2.80
4ofy	Crystal structure of the complex of syg-1 d1-d2 and syg-2 d1-d4	x-ray diffraction	3.30
3l5x	Crystal structure of the complex between il-13 and h2l6 fab	x-ray diffraction	1.90
1k8r	Crystal structure of ras-bry2rbd complex	x-ray diffraction	3.00
3u82	Binding of herpes simplex virus glycoprotein d to	x-ray diffraction	3.16

	nectin-1 exploits host cell adhesion		
5cyk	Structure of ytm1 bound to the c-terminal domain of erb1-r486e	x-ray diffraction	3.00
3lzf	Crystal structure of fab 2d1 in complex with the 1918 influenza virus hemagglutinin	x-ray diffraction	2.80
3vr6	Crystal structure of amp-pnp bound enterococcus hirae v1-atpase [bv1]	x-ray diffraction	2.68
1bd2	Complex between human t-cell receptor b7, viral peptide (tax) and mhc class i molecule hla-a 0201	x-ray diffraction	2.50
2pcc	Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c	x-ray diffraction	2.30
1gl0	Structure of the complex between bovine alpha-chymotrypsin and pmp-d2v, an inhibitor from the insect locusta migratoria	x-ray diffraction	3.00
1gl1	Structure of the complex between bovine alpha-chymotrypsin and pmp-c, an inhibitor from the insect locusta migratoria	x-ray diffraction	2.10
3mzg	Crystal structure of a human prolactin receptor antagonist in complex with the extracellular domain of the human prolactin receptor	x-ray diffraction	2.10
5xco	Crystal structure of human k-ras g12d mutant in complex with gdp and cyclic inhibitory peptide	x-ray diffraction	1.25
4gxu	Crystal structure of antibody 1f1 bound to the 1918 influenza hemagglutinin	x-ray diffraction	3.29
3sgb	Structure of the complex of streptomyces griseus protease b and the third domain of the turkey ovomucoid inhibitor at 1.8 angstroms resolution	x-ray diffraction	1.80
3s9d	Binary complex between ifna2 and ifnar2	x-ray diffraction	2.00
1vfb	Bound water molecules and conformational stabilization help mediate an antigen-antibody association	x-ray diffraction	1.80
1jtg	Crystal structure of tem-1 beta-lactamase / beta-lactamase inhibitor protein complex	x-ray diffraction	1.73
1he8	Ras g12v - pi 3-kinase gamma complex	x-ray diffraction	3.00
4k71	Crystal structure of a high affinity human serum albumin variant bound to the neonatal fc receptor	x-ray diffraction	2.40
3kud	Complex of ras-gdp with rafrbd(a85k)	x-ray diffraction	2.15
1xxm	The modular architecture of protein-protein binding site	x-ray diffraction	1.90
2nyy	Crystal structure of botulinum neurotoxin type a complexed with monoclonal antibody cr1	x-ray diffraction	2.61
1r0r	1.1 angstrom resolution structure of the complex between the protein inhibitor, omtky3, and the serine protease, subtilisin carlsberg	x-ray diffraction	1.10
5tar	Crystal structure of farnesylated and methylated kras4b in complex with pde-delta (crystal form ii - with ordered hypervariable region)	x-ray diffraction	1.90
3se4	Human ifnw-ifnar ternary complex	x-ray diffraction	3.50

3se3	Human ifna2-ifnar ternary complex	x-ray diffraction	4.00
2kso	Epha2:ship2 sam:sam complex	NMR	NA
1ppf	X-ray crystal structure of the complex of human leukocyte elastase (pmn elastase) and the third domain of the turkey ovomucoid inhibitor	x-ray diffraction	1.80
4krl	Nanobody/vhh domain 7d12 in complex with domain iii of the extracellular region of egfr, ph 6.0	x-ray diffraction	2.85
4kro	Nanobody/vhh domain ega1 in complex with the extracellular region of egfr	x-ray diffraction	3.05
3eg5	Crystal structure of mdia1-tsh gbd-fh3 in complex with cdc42-gmppnp	x-ray diffraction	2.70
2p5e	Crystal structures of high affinity human t-cell receptors bound to pmhc reveal native diagonal binding geometry	x-ray diffraction	1.89
1ahw	A complex of extracellular domain of tissue factor with an inhibitory fab (5g9)	x-ray diffraction	3.00
2ny7	Hiv-1 gp120 envelope glycoprotein complexed with the broadly neutralizing cd4-binding-site antibody b12	x-ray diffraction	2.30
2g2u	Crystal structure of the shv-1 beta-lactamase/beta-lactamase inhibitor protein (blip) complex	x-ray diffraction	1.60
1fcc	Crystal structure of the c2 fragment of streptococcal protein g in complex with the fc domain of human igg	x-ray diffraction	3.20
1brs	Protein-protein recognition: crystal structural analysis of a barnase- barstar complex at 2.0-a resolution	x-ray diffraction	2.00
1a22	Human growth hormone bound to single receptor	x-ray diffraction	2.60
2vn5	The clostridium cellulolyticum dockerin displays a dual binding mode for its cohesin partner	x-ray diffraction	1.90
1bj1	Vascular endothelial growth factor in complex with a neutralizing antibody	x-ray diffraction	2.40
3ngb	Crystal structure of broadly and potently neutralizing antibody vrc01 in complex with hiv-1 gp120	x-ray diffraction	2.68
5e9d	Rd-1 mart-1 high bound to mart-1 decameric peptide (ela) in complex with hla-a2	x-ray diffraction	2.51
4yh7	Crystal structure of ptpdelta ectodomain in complex with il1rapl1	x-ray diffraction	4.40
1jrh	Complex (antibody/antigen)	x-ray diffraction	2.80
1c4z	Structure of an e6ap-ubch7 complex: insights into the ubiquitination pathway	x-ray diffraction	2.60
4uyp	High resolution structure of the third cohesin scac in complex with the scab dockerin with a mutation in the n-terminal helix (in to si) from acetivibrio cellulolyticus displaying a type i interaction.	x-ray diffraction	1.49
1fss	Acetylcholinesterase (e.c. 3.1.1.7) complexed with fasciculin-ii	x-ray diffraction	3.00
5c6t	Crystal structure of hcmv glycoprotein b in	x-ray diffraction	3.60

	complex with 1g2 fab		
4gu0	Crystal structure of lsd2 with h3	x-ray diffraction	3.10
3uii	Crystal structure of human survivin in complex with h3(1-10) peptide	x-ray diffraction	2.60
3aaa	Crystal structure of actin capping protein in complex with v-1	x-ray diffraction	2.20
1b41	Human acetylcholinesterase complexed with fasciculin-ii, glycosylated protein	x-ray diffraction	2.76
4nkq	Structure of a cytokine receptor complex	x-ray diffraction	3.30
1cho	Crystal and molecular structures of the complex of alpha-*chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 angstroms resolution	x-ray diffraction	1.80
2wpt	The crystal structure of im2 in complex with colicin e9 dnase	x-ray diffraction	1.78
2nz9	Crystal structure of botulinum neurotoxin type a complexed with monoclonal antibody ar2	x-ray diffraction	3.79
4wnd	Crystal structure of the tpr domain of lgn in complex with frmpd4/preso1 at 1.5 angstrom resolution	x-ray diffraction	1.50
3pwp	The complex between tcr a6 and human class i mhc hla-a2 with the bound hud peptide	x-ray diffraction	2.69
4j2l	Crystal structure of axh domain complexed with capicua	x-ray diffraction	3.15
1b2u	Structural response to mutation at a protein-protein interface	x-ray diffraction	2.10
1b2s	Structural response to mutation at a protein-protein interface	x-ray diffraction	1.82
1ao7	Complex between human t-cell receptor, viral peptide (tax), and hla-a 0201	x-ray diffraction	2.60
1dvf	Idiotopic antibody d1.3 fv fragment-antiidiotopic antibody e5.2 fv fragment complex	x-ray diffraction	1.90
2ccl	The s45a, t46a mutant of the type i cohesin-dockerin complex from the cellulosome of clostridium thermocellum	x-ray diffraction	2.03
1bp3	The xray structure of a growth hormone-prolactin receptor complex	x-ray diffraction	2.90
2qj9	Crystal structure analysis of bmp-2 in complex with bmpr-ia variant b1	x-ray diffraction	2.44
2noj	Crystal structure of ehp / c3d complex	x-ray diffraction	2.70
5ufe	Wild-type k-ras(gnp)/r11.1.6 complex	x-ray diffraction	2.30
1kbh	Mutual synergistic folding in the interaction between nuclear receptor coactivators cbp and actr	NMR	NA
3hfm	Structure of an antibody-antigen complex. crystal structure of the hy/hel-10 fab-lysozyme complex	x-ray diffraction	3.00
3qdg	The complex between tcr dmf5 and human class i mhc hla-a2 with the bound mart-1(26-35)(a27l) peptide	x-ray diffraction	2.69

3qdj	The complex between tcr dmf5 and human class i mhc hla-a2 with the bound mart-1(27-35) nonameric peptide	x-ray diffraction	2.30
4yfd	Crystal structure ptp delta ig1-fn2 in complex with il-1racp	x-ray diffraction	3.25
2j0t	Crystal structure of the catalytic domain of mmp-1 in complex with the inhibitory domain of timp-1	x-ray diffraction	2.54
5ufq	K-rasg12d(gnp)/r11.1.6 complex	x-ray diffraction	2.20
3mzw	Her2 extracelluar region with affinity matured 3-helix affibody zher2:342	x-ray diffraction	2.90
3g6d	Crystal structure of the complex between cnto607 fab and il-13	x-ray diffraction	3.20
4x4m	Structure of fcgammari in complex with fc reveals the importance of glycan recognition for high affinity igg binding	x-ray diffraction	3.49
3m63	Crystal structure of ufd2 in complex with the ubiquitin-like (ubl) domain of dsk2	x-ray diffraction	2.40
4myw	Structure of hsv-2 gd bound to nectin-1	x-ray diffraction	3.19
4e6k	2.0 a resolution structure of pseudomonas aeruginosa bacterioferritin (bfrb) in complex with bacterioferritin associated ferredoxin (bfd)	x-ray diffraction	2.00
1b3s	Structural response to mutation at a protein-protein interface	x-ray diffraction	2.39
1z7x	X-ray structure of human ribonuclease inhibitor complexed with ribonuclease i	x-ray diffraction	1.95
4jpk	Crystal structure of the germline-targeting hiv-1 gp120 engineered outer domain eod-gt6 in complex with a putative vrc01 germline precursor fab	x-ray diffraction	2.40
4rs1	Crystal structure of receptor-cytokine complex	x-ray diffraction	2.68
1emv	Crystal structure of colicin e9 dnase domain with its cognate immunity protein im9 (1.7 angstroms)	x-ray diffraction	1.70
1sbb	T-cell receptor beta chain complexed with superantigen seb	x-ray diffraction	2.40
1mlc	Monoclonal antibody fab d44.1 raised against chicken egg- white lysozyme complexed with lysozyme	x-ray diffraction	2.50
1ohz	Cohesin-dockerin complex from the cellulosome of clostridium thermocellum	x-ray diffraction	2.20
2abz	Crystal structure of c19a/c43a mutant of leech carboxypeptidase inhibitor in complex with bovine carboxypeptidase a	x-ray diffraction	2.16
4cvw	Structure of the barley limit dextrinase-limit dextrinase inhibitor complex	x-ray diffraction	2.67
4y61	Crystal structure of the complex between slitrk2 lrr1 and ptp delta ig1-fn1	x-ray diffraction	3.36
4g0n	Crystal structure of wt h-ras-gppnhp bound to the rbd of raf kinase	x-ray diffraction	2.45
2qja	Crystal structure analysis of bmp-2 in complex with bmpr-ia variant b12	x-ray diffraction	2.60

1wqj	Structural basis for the regulation of insulin-like growth factors (igfs) by igf binding proteins (igfbps)	x-ray diffraction	1.60
2qjb	Crystal structure analysis of bmp-2 in complex with bmpr-ia variant ia/ib	x-ray diffraction	2.50
4krp	Nanobody/vhh domain 9g8 in complex with the extracellular region of egfr	x-ray diffraction	2.82
3bp8	Crystal structure of mlc/eiib complex	x-ray diffraction	2.85

Table S2 - Distribution of multiple point mutations across different experimental methods available in SKEMPI2 and used in this work.

Acronym	**Technique**	**# mutations**
SPR	Surface Plasmon Resonance	599
FL	Fluorescence	355
SFFL	Stopped Flow Fluorescence	170
ITC	Isothermal Titration	138
SP	Spectroscopy	115
IASP	Spectroscopy Inhibition Assay	81
ELISA	ELISA	70
RA	Radioactive Ligand Binding	63
IAFL	Fluorescence Inhibition Assay	54
KinExA	Kinetic Exclusion Assay	47
IARA	Radioligand Inhibition Assay	11
ELFA	Enzyme-linked Functional Assay	10
BI	Biolayer Interferometry	3
Other	SE, IAGE and ESMA	3
SPR,SFFL	SPR,SFFL	2

Table S3 - Complementary features used to model the effects of multiple point mutations on PPIs.

Category	**Description**	**Tool**
Normal Mode Analysis	Deformation energy and atomic fluctuation across 4 different force-fields (C-alpha, ANM, pfANM, REACH, sdENM)	Bio3D (1)
Residue Environment	Torsion angles (psi and phi), relative solvent accessibility and residue depth	Biopython (2)
Evolutionary and contact potential	Scores from substitution tables	AAINDEX (3), Blosum and PAM matrices
Non-covalent contacts	Hydrogen bonds, Hydrophobic contacts, PI stacking and Ionic interactions	Arpeggio (4)

Wild-type inter-residue distance	Average, shortest and longest distances among wild-type residues being mutated	Python
Individual $\Delta\Delta G^{binding}$	Calculated for each single-point mutation separately	mCSM-PPI2 (5)

Table S4 - Performance of mmCSM-PPI for different supervised learning algorithms. Evaluation metrics were calculated using 10-fold cross validation before feature selection.

Algorithm	**Pearson**	**Kendall**	**Spearman**	**RMSE (kcal/mol)**
Extra Trees	0.73	0.53	0.73	1.66
Random Forest	0.70	0.52	0.72	1.75
Gradient Boosting	0.69	0.50	0.70	1.79
XGBoost	0.68	0.49	0.68	1.81

Table S5 - Parameters used in the final predictive model for mmCSM-PPI using Extra Trees algorithm available in the Scikit-learn Python library.

Hyperparameter	**Values**
n_estimators	300
min_samples_split	10
min_samples_leaf	3
max_depth	40
bootstrap	False

Table S6 - Feature importance from the Extra Trees algorithm used for the final model of mmCSM-PPI.

Feature	**Score**
Sum of $\Delta\Delta G^{binding}$ for each single point mutation separately (mCSM-PPI2)	0.480
Average graph-based signatures of wild-type residues	0.462
Sum of scores from DOSZ010103 (AAINDEX)	0.015
Average Pharmacophore changes (Positives)	0.007
Average deformation energies of wild-type residues (Bio3D)	0.006
Sum of Pharmacophore changes (Sulfurs)	0.005
Sum of scores from RUSR970103 (AAINDEX)	0.004
Sum of deformation energies of wild-type residues (Bio3D)	0.004
Sum of Hydrophobic contacts of wild-type residues (Arpeggio)	0.003
Average atomic fluctuation of wild-type residues (Bio3D)	0.003
Average Weak Polar interactions of wild-type residues (Arpeggio)	0.003
Sum of scores for MEHP950101 (AAINDEX)	0.003

Shortest distance between wild-type residues	0.003
Sum of Pharmacophore changes (Negatives)	0.002
Average Phi torsion angle of wild-type residues	0.002

Table S7 - Performance comparison of mmCSM-PPI and predictive models using only the most important features on a non-redundant blind test comprising entries with 4 or more mutations.

	Pearson	Kendall	Spearman	RMSE
mmCSM-PPI	0.70	0.48	0.64	2.06
Sum of Individual ΔΔGbinding	0.61*	0.41#	0.57+	2.55a
Graph-based signatures	0.37*	0.26#	0.37+	2.83a

* p-value < 0.05 by Fisher r-to-z transformation test
# p-value < 0.05 by transforming tau-to-r followed by Fisher r-to-z transformation
+ p-value < 0.05 by transforming rho-to-r followed by Fisher r-to-z transformation
a p-value < 0.05 by Diebold-Mariano test

Table S8- Distribution of multiple mutations across different protein-protein complex structures in the PDB extracted from SKEMPI2.

Protein-protein complex (PDB)	**# multiple mutations**
1JTG	136
3SGB	88
1CHO	84
1KBH	83
1R0R	76
2B2X	67
3S9D	61
1PPF, 1AO7	56
1BRS	45
4G0N	42
1A22	37
3L5X	34
1KNE	33
3SE3	32
2C5D, 1LFD	31
1DAN	29
1REW	25

2VN5	24
3MZW	22
1MHP	20
2G2U, 5XCO	18
2WPT	17
4NKQ, 3HFM, 1MLC, 1CZ8	16
3KUD	15
1HE8, 1VFB, 1BP3	14
1EMV, 1DQJ, 1A4Y, 1DVF	13
3VR6, 4MNQ, 1JRH, 1BJ1	12
3IDX, 2NY7	11
1QAB, 4K71, 4FTV	9
3HG1, 4RS1, 4UYP, 4UYQ, 5E9D	8
4RA0, 1AHW, 1MQ8, 5M2O, 1BD2	7
3BE1, 3G6D, 2PCC, 2J0T, 4L3E	6
2KSO, 2DVW, 2PYE, 4JFE, 4JFD, 1OHZ, 3BDY, 1YCS	5
2OOB, 1FSS, 1Z7X, 1TM1, 2P5E, 1MAH, 3NGB	4
2QJA, 5UFE, 1YY9, 4GNK, 2QJ9, 5CYK, 2QJB, 4YH7, 4E6K, 1GUA, 4J2L, 1C1Y, 3D3V, 4B0M	3
3EG5, 1GC1, 1YQV, 2NYY, 4OFY, 1B41, 1N8Z, 3QIB, 5UFQ, 3QDG, 3QDJ, 4YFD, 3SE4, 5C6T	2
2ABZ, 4KRL, 1FCC, 1E50, 1XXM, 3PWP, 1C4Z, 3BP8, 4JEU, 1EFN, 1AK4, 4MYW, 4CVW, 4X4M, 1WQJ, 4GXU, 2NOJ, 1B2S, 2NZ9, 3AAA, 2O3B, 1B2U, 4KRP, 4Y61, 4KRO, 4WND, 3UIH, 1GL0, 1GL1, 4G2V, 3LZF, 5TAR, 3U82, 3MZG, 1K8R, 1Y4A, 4JPK, 4GU0, 3UII, 3M63, 2CCL, 1B3S, 1SBB	1

Table S9 - Performance comparison on increasing and decreasing mutations.

	Decreasing affinity				**Increasing affinity**			
Method	Pearson	Kendall	Spearman	RMSE	Pearson	Kendall	Spearman	RMSE
mmCSM-PPI	0.72	0.46	0.64	1.67	0.16	0.21	0.31	2.93
Discovery Studio	0.30[*]	0.30[#]	0.44[+]	4.84[a]	0.18	0.11	0.17	5.42
FoldX	0.34[*]	0.21[#]	0.32[+]	2.83[a]	0.23	0.20	0.32	2.83

[*] p-value < 0.05 by Fisher r-to-z transformation test

[#] p-value < 0.05 by transforming tau-to-r followed by Fisher r-to-z transformation

[+] p-value < 0.05 by transforming rho-to-r followed by Fisher r-to-z transformation

[a] p-value < 0.05 by Diebold-Mariano test

Table S10 - mmCSM-PPI classification by regression using different thresholds.

		$\|\Delta\Delta G^{binding}\| > 0.5$				$\|\Delta\Delta G^{binding}\| > 1.0$				$\|\Delta\Delta G^{binding}\| > 1.5$			
Threshold	**Class**	**Precision**	**Recall**	**MCC**	**AUC**	**Precision**	**Recall**	**MCC**	**AUC**	**Precision**	**Recall**	**MCC**	**AUC**
$\Delta\Delta G^{mmCSM-PPI} > -1.0$	Increase	0.46	0.76	0.44	0.76	0.45	0.80	0.47	0.78	0.45	0.79	0.48	0.79
$\Delta\Delta G^{mmCSM-PPI} < -1.0$	Decrease	0.92	0.75	0.44	0.76	0.94	0.77	0.47	0.78	0.95	0.80	0.48	0.79
$\Delta\Delta G^{mmCSM-PPI} > -0.50$	Increase	0.52	0.61	0.43	0.73	0.53	0.65	0.48	0.76	0.52	0.65	0.48	0.76
$\Delta\Delta G^{mmCSM-PPI} < -0.50$	Decrease	0.88	0.84	0.43	0.73	0.91	0.86	0.48	0.76	0.92	0.87	0.48	0.76
$\Delta\Delta G^{mmCSM-PPI} > 0$	Increase	0.76	0.43	0.49	0.70	**0.74**	**0.49**	**0.53**	**0.72**	0.77	0.48	0.55	0.73
$\Delta\Delta G^{mmCSM-PPI} < 0$	Decrease	0.86	0.96	0.49	0.70	**0.89**	**0.96**	**0.53**	**0.72**	0.90	0.97	0.55	0.73
$\Delta\Delta G^{mmCSM-PPI} > 0.50$	Increase	0.94	0.29	0.46	0.64	0.94	0.36	0.53	0.68	0.96	0.39	0.57	0.69
$\Delta\Delta G^{mmCSM-PPI} < 0.50$	Decrease	0.83	0.99	0.46	0.64	0.87	0.99	0.53	0.68	0.8	1.00	0.57	0.69
$\Delta\Delta G^{mmCSM-PPI} > 1.00$	Increase	1.00	0.18	0.38	0.59	1.00	0.24	0.45	0.62	1.00	0.24	0.46	0.62
$\Delta\Delta G^{mmCSM-PPI} < 1.00$	Decrease	0.81	1.00	0.38	0.59	0.84	1.00	0.45	0.62	0.86	1.00	0.46	0.62

Table S11 - Performance comparison for blind-test non-redundant at the mutation level.

Method	Pearson	Kendall	Spearman	RMSE (kcal/mol)
mmCSM-PPI	0.67	0.47	0.67	1.72
Discovery Studio	0.36*	0.28#	0.38+	2.74a
FoldX	0.29*	0.26#	0.39+	4.55a

* p-value < 0.05 by Fisher r-to-z transformation test

p-value < 0.05 by transforming tau-to-r followed by Fisher r-to-z transformation

+ p-value < 0.05 by transforming rho-to-r followed by Fisher r-to-z transformation

a p-value < 0.05 by Diebold-Mariano test

FIGURES

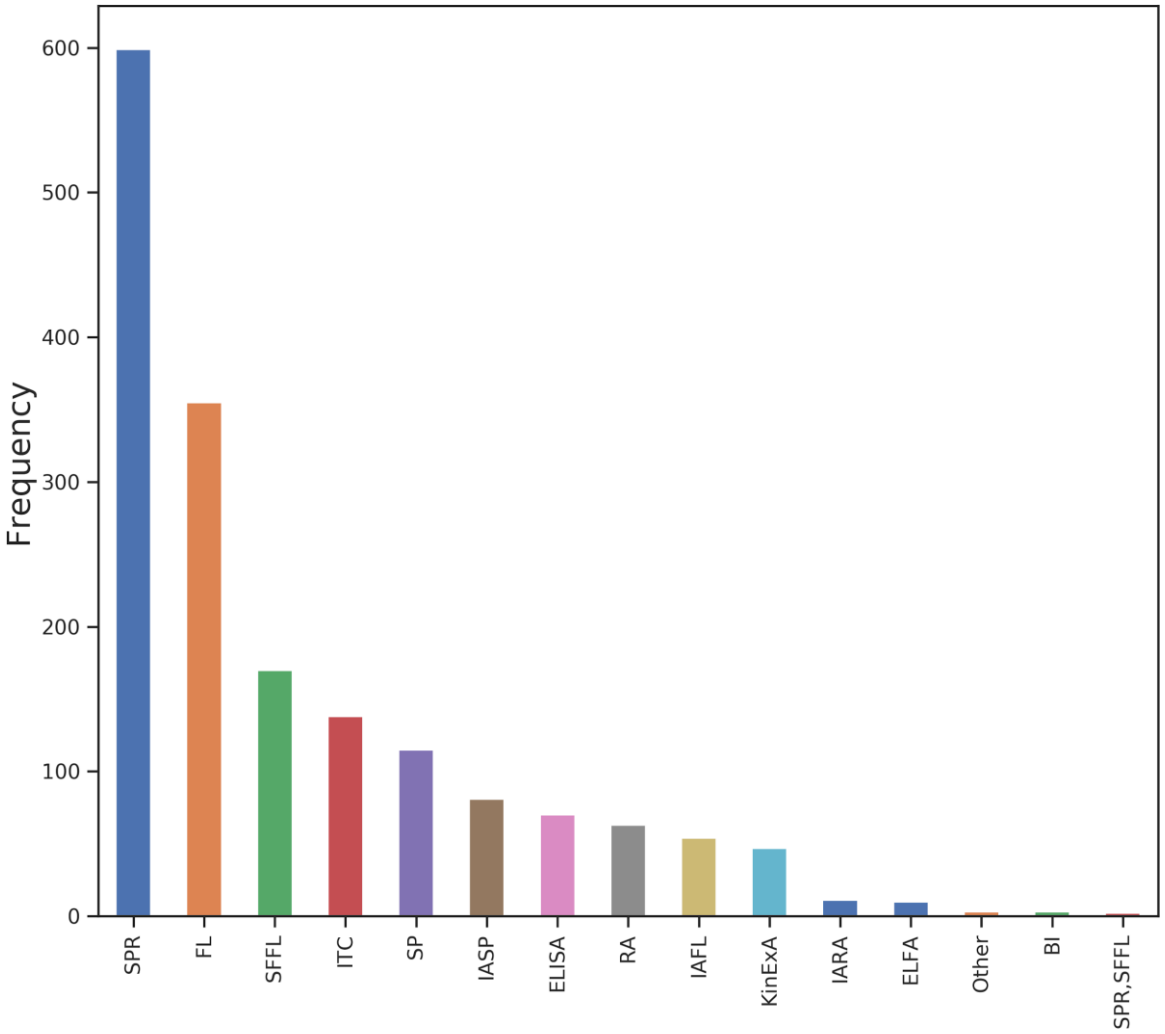


Figure S1 - Distribution of experimental techniques used to generate the data set extracted from SKEMPI2. 5 experimental techniques represent the majority of the dataset (80%): surface plasmon resonance (SPR), fluorescence (FL), stopped flow fluorescence (SFFL), isothermal titration calorimetry (ITC) and spectroscopy.

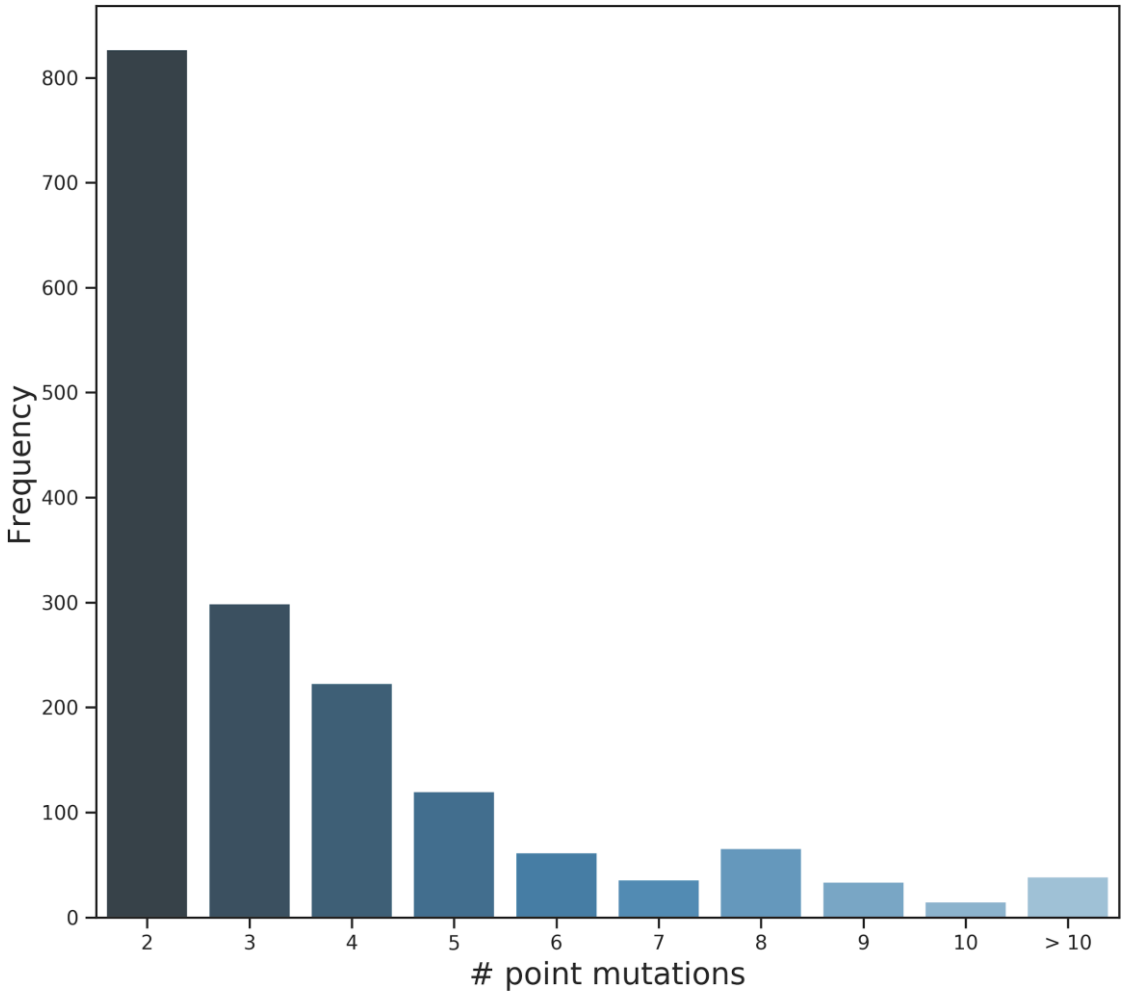


Figure S2 - Distribution of multiple mutations across the data retrieved from SKEMPI2. Double and triple mutants, which account for more than 65% (1126) of all entries, were used for training mmCSM-PPI. The remaining 595 entries were held out and used as a non-redundant blind test at mutation level.

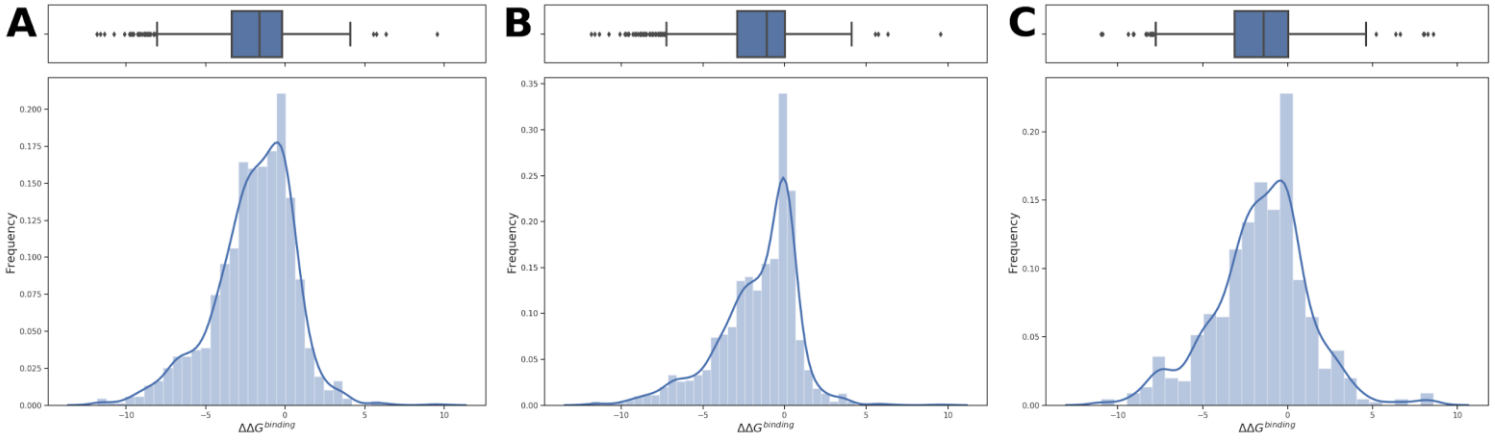


Figure S3 - $\Delta\Delta G^{binding}$ distribution of training and blind test sets used in this work. A) depicts the distribution of changes in binding affinity for the original training set extracted from SKEMPI2. B) shows the distribution on the training set after including hypothetical reverse mutations. Finally, C) summarises the $\Delta\Delta G^{binding}$ on the non-redundant blind test set used for validation and comparisons with other methods.

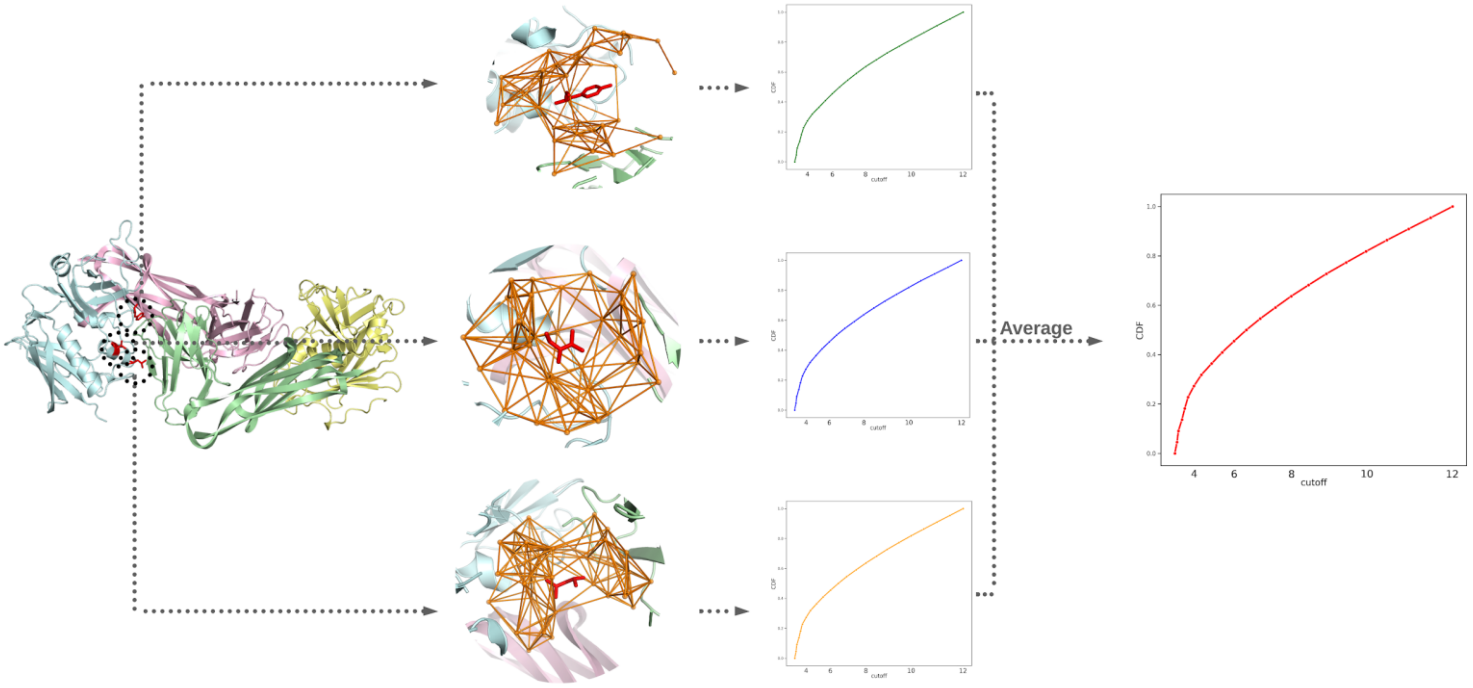


Figure S4 - Graph-based signatures representation for multiple point mutations. For each point mutation, the residue environment is represented as a graph where surrounding residues are represented as nodes and their interactions as edges. Distance patterns between atoms characterised by their properties are compiled as cumulative distributions for each environment separately. Finally, the distributions are averaged based on the number of point mutations.

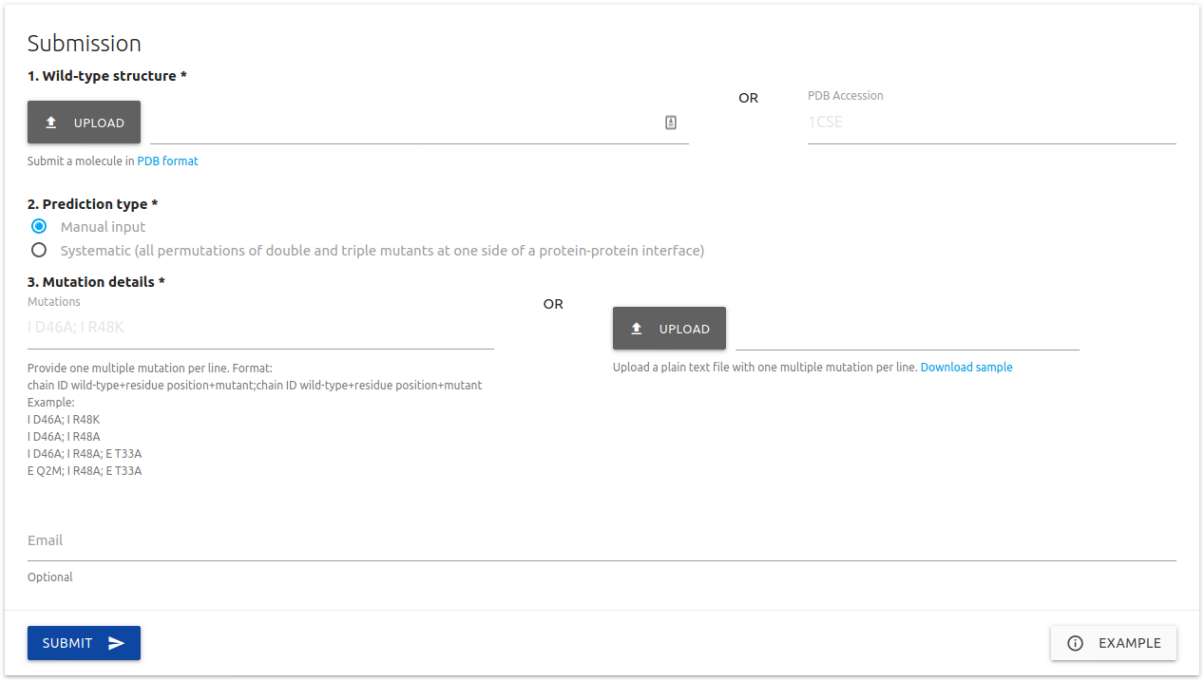


Figure S5 - mmCSM-PPI submission page. Two predictive types are available: Manual input and Systematic evaluation. In both cases, users must provide the 3D structure of a protein complex by either uploading a file in PDB format or specifying a valid PDB accession code. For the Manual input option, a list of mutations of interest are required, represented as the chain identifier, the wild-type residue one-letter code, the residue number and the mutant residue one-letter code, and with each point mutation separated by a semi-colon. For the Systematic evaluation option, users must specify a chain identifier from which interfaces are automatically identified and the effects of all double and triple point mutations are assessed.

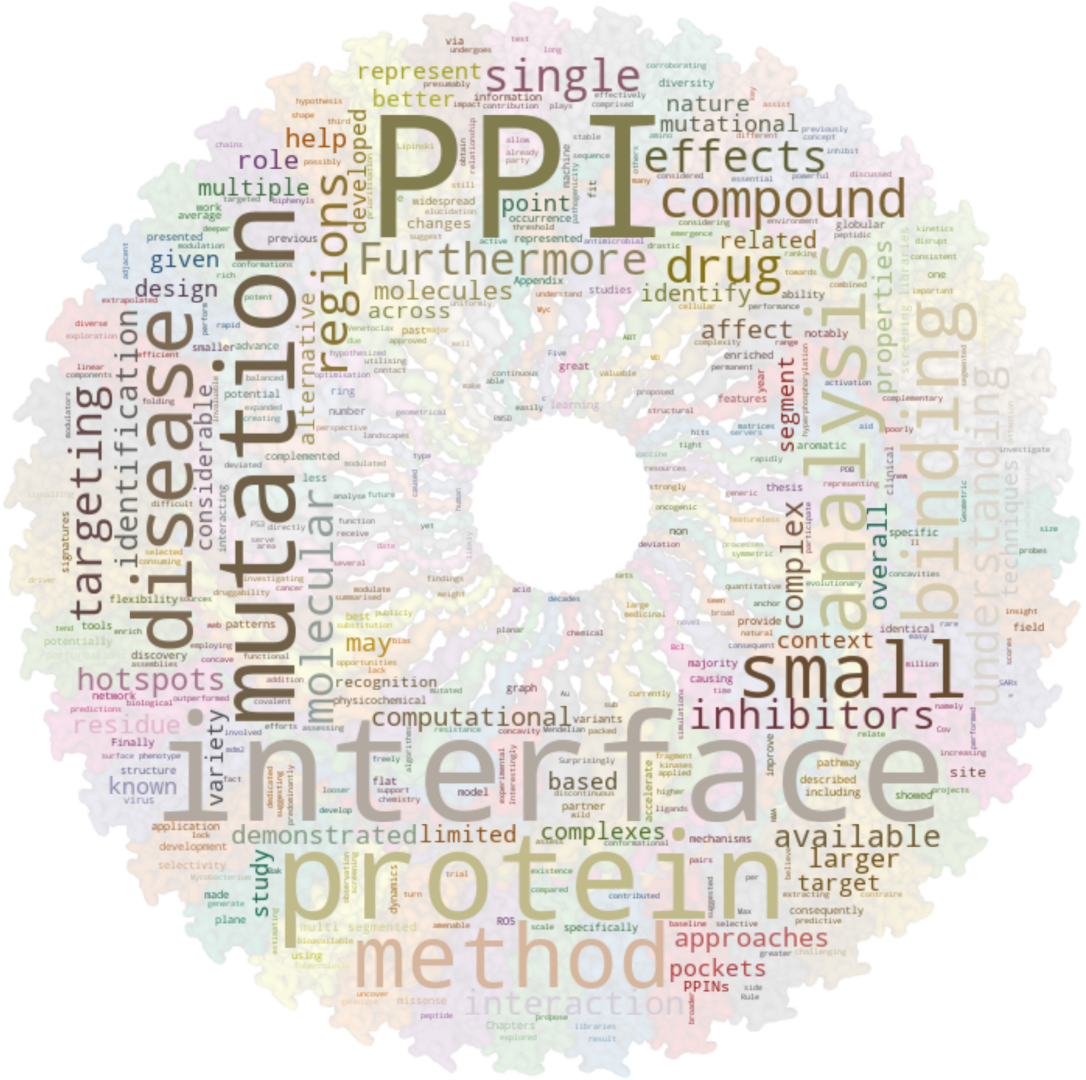
Figure S6 - mmCSM-PPI results page. For both predictive options, manual input and systematic evaluation, results are shown as a downloadable table where predicted effects of multiple mutations in $\Delta\Delta G^{binding}$ and individual predicted effects for each single-point mutation are also available. In addition, an interactive 3D viewer is also available where interactions for an entry are displayed. Users can change the interactions in the interactive viewer by selecting entries from the table.

REFERENCES

1. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695-2696. http://dx.doi.org/10.1093/bioinformatics/btl461
2. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422-1423. http://dx.doi.org/10.1093/bioinformatics/btp163
3. Kawashima, S. and Kanehisa, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*, **28**, 374. http://dx.doi.org/10.1093/nar/28.1.374
4. Jubb, H.C., Higueruelo, A.P., Ochoa-Montano, B., Pitt, W.R., Ascher, D.B. and Blundell, T.L. (2017) Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J Mol Biol*, **429**, 365-371. http://dx.doi.org/10.1016/j.jmb.2016.12.004
5. Rodrigues, C.H.M., Myung, Y., Pires, D.E.V. and Ascher, D.B. (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res*, **47**, W338-W344. http://dx.doi.org/10.1093/nar/gkz383

Chapter 8

Discussion and Conclusions



For the past decades, considerable attention has been given to PPIs, particularly aiming to target and modulate them, given their selectivity and their widespread roles in cellular processes. However, the design of drugs to target PPIs has been long considered challenging, given the complex and diverse nature of these regions. In this context, the identification and study of hotspots residues, and the emergence of publicly available libraries of compounds with experimental information to serve as probes, represent considerable advances to support and accelerate drug discovery efforts.

In this work, I hypothesized that computational techniques can provide powerful and valuable biological insight into the nature of PPIs and how they can be effectively modulated. In Chapter 2, I performed a large-scale analysis of the structural and chemical properties of all PPI interfaces available on the PDB, and demonstrated that PPI interfaces are not uniformly flat and featureless as has been previously described. In fact, quantitative analysis of PPI use of concavity showed that while on average these regions are flat, the majority of interfaces made use of small concavities, corroborating the anchor residues hypothesis and the widespread occurrence of complemented pockets.

Furthermore, multi segmented interfaces utilising discontinuous binding regions were not only larger than single segmented interfaces, but also less well packed and less complementary in shape. These findings suggest that single interacting segments make tight selective interactions with their globular partner, compared to looser interfaces in larger multi segmented complexes. Surprisingly, single segment interfaces have shown to be more planar than multi segmented interfaces. This observation may be related to the potentially linear nature of peptide binding sites, binding across a small interface surface area. While peptidic interfaces were overall more concave, their deviation from a best-fit plane through the interface may have been smaller due to smaller interface size. *Au contraire*, interfaces for non-identical and identical symmetric pairs were larger and, on average, comprised multiple interacting segments, consequently increasing the RMSD from an interface best-fit plane. From a druggability perspective, continuous binding sites that make use of deeper binding pockets and significantly (ANOVA p-value <0.5) more directional non-covalent interactions (hydrogen/polar bonds) may be more immediately amenable for targeting than discontinuous interfaces. In addition, the presence of completed sub-pockets across most globular interfaces may be explored by small-molecule fragments.

As discussed in Chapter 1, PPI interface regions are known to be enriched with disease-mutations, and the ability to rapidly uncover the different mechanisms by which these variants disrupt protein interactions is essential to help identify disease driver mutations and hotspots, design of more stable complexes, and better understand disease development. In Chapters 3, 4, 6 and 7, I have demonstrated that protein structure and sequence can be used as rich sources of information from which computational approaches can be developed. By employing machine learning techniques (Appendix A), I developed methods that contributed for a better understanding of how missense mutations affect PPIs. To this date, the majority of methods currently available for assessing the effects of missense mutations on PPIs are known to perform poorly in new test sets, and are limited to analysis of the effects of single point mutations. I have demonstrated in Chapters 3, 4 and 6 that my approaches outperformed many other alternative methods on estimating these effects of single point mutations with more balanced predictions. Furthermore, I extrapolated these approaches to analyse mutational effects on a multiple mutations context (Chapters 3 and 7), which represented a considerable advance to the field.

Over the past 3 years, these mutational analysis methods have been applied into a variety of projects, most notably the exploration of the mutation landscapes across all proteins from the SARs-Cov-2 virus [89] and several other studies investigating this rapid mutated virus [90–93]; antimicrobial drug resistance and pathogenicity (*i.e.*, *Mycobacterium Tuberculosis*) [94–96]; identification of hotspots at PPI interfaces which could, in turn, be used to aid drug/vaccine design [97]; to investigate the role of mutations in cancer [98, 99] and rare diseases [100, 101]. Overall, these studies showed that up to 60% of mutations involved in drug resistance and disease phenotype would lead to significant disruption of key protein-protein interactions. The findings described in Chapters 3, 4, 6 and 7 have paved the way to a new generation of tools that will greatly benefit our understanding of how mutations affect protein interactions and how we can develop drugs to selectively target them with implications in personalised medicine.

The variety of mechanisms by which mutations may affect protein function was also represented in the diversity of features selected by the machine learning algorithms for these predictive models. Overall, features representing geometrical and physicochemical properties of wild-type residue environment (graph-based signatures), protein dynamics, contact potentials and evolutionary scores from substitution matrices had greater impact in method performance. Most notably, I demonstrated that extracting protein flexibility

properties using NMA is a great alternative to computational and time consuming MD simulations.

Protein dynamics plays an important role in molecular recognition with the occurrence of PPIs where one one of the partners undergoes small or more drastic conformational changes. Changes in protein flexibility, caused by point mutations, can directly affect binding site regions involved in protein recognition, or even "lock" a protein in specific conformations, as seen for mutations causing permanent activation of protein kinases and consequent hyperphosphorylation of their targets. Geometric and physicochemical changes relate to the ability of amino acid side chains to participate in non-covalent interactions which are key components for the protein folding, kinetics of PPI binding and also molecular recognition.

Finally, as we improve our understanding of the molecular properties at PPI interfaces, a number of PPI inhibitors/modulators have been proposed, some of which are already approved for clinical use, namely Venetoclax (ABT-199) [102] and others under clinical trial. However, the complexity and diversity of PPI interfaces combined with the lack of screening libraries dedicated specifically to PPIs, or yet limited to compounds targeting a small number of protein complexes, still makes it difficult to identify potential inhibitors. In Chapter 5, I showed that most PPI inhibitors used deviate from the Lipinski Rule of Five (RO5) for bioavailable drugs, suggesting a natural bias in the screening libraries used to obtain these molecules. Interestingly, most active compounds presented a larger molecular weight than the threshold used in RO5, presumably the result of compound optimisation via medicinal chemistry to generate small molecules with a higher selectivity towards specific PPIs. Furthermore, more potent compounds were enriched with complex ring structures, including biphenyls, consistent with previous studies demonstrating that PPI inhibitors tend to have more aromatic rings than other ligands, and possibly related to hotspots targeting as these are known to be predominantly aromatic residues. Furthermore, I explored the concept of graph-based signatures to model small molecules inhibitors for targeting 3 oncogenic PPIs: Bcl2/Bak, mdm2/P53 and c-Myc/Max. Using these as a baseline I was able to develop a generic method that represents a major contribution into compound prioritisation/ranking and potentially as an alternative to enrich screening libraries.

In future work, I would like to further assess the effects of mutations in the context

of protein-protein interaction networks (PPINs), which could help in the identification of adjacent proteins in the same network (*i.e.*, same signalling pathway) that could be more easily targeted for modulation. In addition, a broader mutational analysis of PPINs could improve our understanding of how multiple variants affect more complex protein assemblies and pathways. Furthermore, a previous study [37] has suggested the existence of perturbations patterns on PPINs across a variety of human Mendelian diseases, and propose a characterisation of molecular interactions to help identify disease-causing mutations. The application of the analysis and computational techniques presented in this thesis can provide a deeper elucidation of how these perturbations patterns are related to disease and consequently allow for a better understanding of the complex genotype-phenotype relationship.

All methods/tools developed and described in this thesis were made freely available via easy-to-use web servers and overall receive over 1 million hits per year. Tools and third-party applications used for the development of these methods are summarised in Appendix B. I strongly believe these represent invaluable resources that can help to accelerate analysis in a broad range of fields, including, but not limited to, protein functional analysis, understanding the role of mutations in diseases and more efficient screening of compounds targeting PPIs.

Appendix A

Machine Learning

A.1 Overview

Machine learning (ML) is a branch of computer science that aims to use historical data to derive a model that describes a phenomenon, aiming for its generalisation (applying it in data never seen before), or to identify inherent patterns within a collection of data. ML has been widely utilised for a variety of tasks such as speech/pattern recognition [103, 104], prediction of phenotype diseases from variants [55] and binding site detections [105]. In that sense, ML can be split into two categories depending on the type of data that is being worked with: supervised and unsupervised learning.

Supervised learning task refers to inferring a mapping function between an input and an output label, using labeled training data. The training data takes the form of a collection of (X,y) pairs in which X is usually represented by a vector of features and y is the known output label (either numerical or categorical) for a given X (Figure A.1A). The goal is to produce a prediction in response to a query with an unknown outcome, based on the information and patterns extracted from the training step [106]. This could also be further divided into classification and regression.The former aims to identify patterns within a dataset that could be used to assign a categorical value (class) to each data point [107], while regression aims to estimate an actual numerical value based on a set of input features.

Alternatively unsupervised learning is defined when the input data comprises a collection of features that describe the dataset without a corresponding label/class associated to

it [108]. In this case, the purpose of ML is to find inherent groupings in the data or to identify relationships, also known as association rules, among the features. The patterns discovered for this type of algorithms commonly have to be manually evaluated or via application of a supervised learning (Figure A.1B)

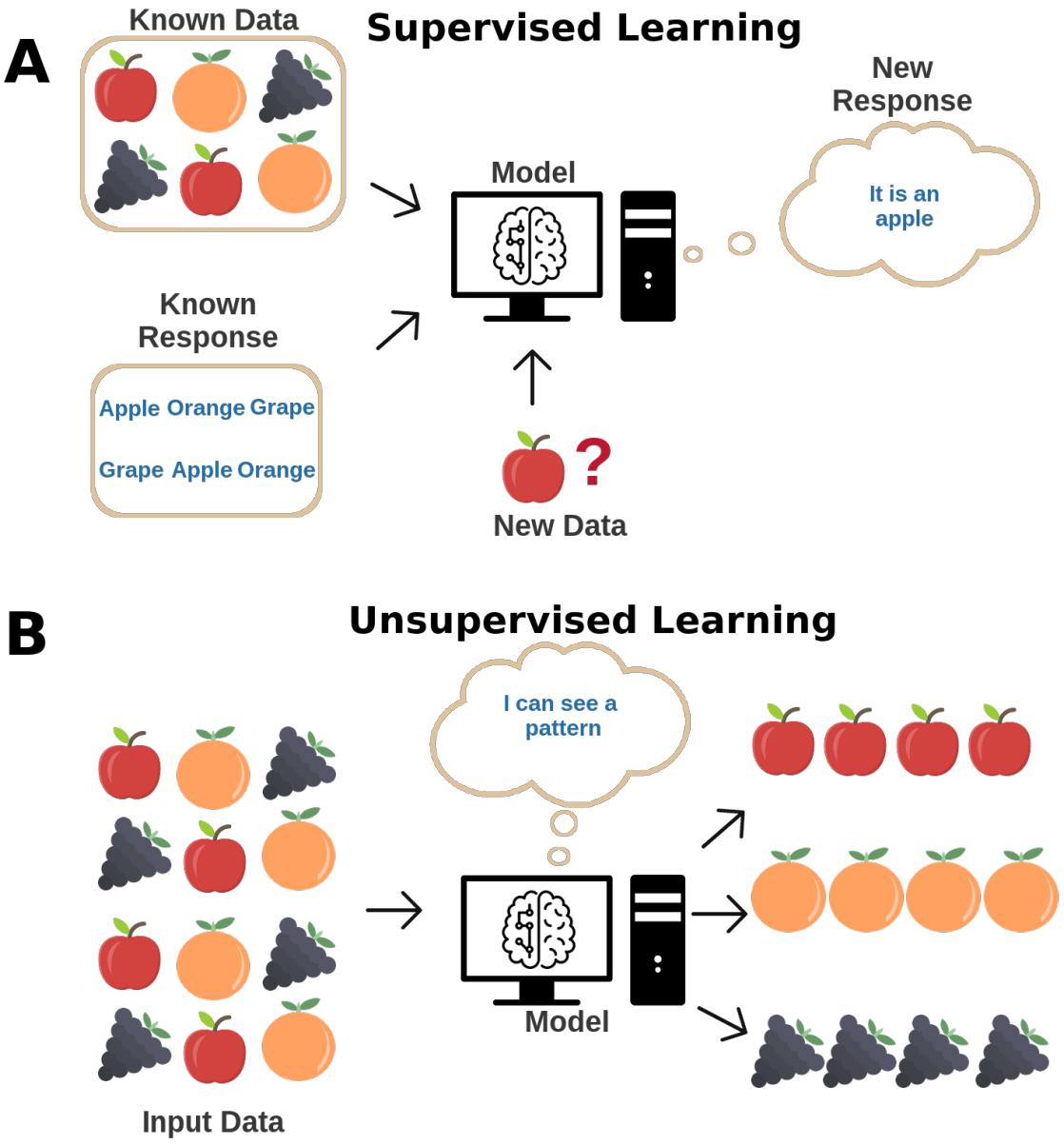


FIGURE A.1: Supervised and Unsupervised learning workflows. ML algorithms for supervised learning (A) are provided with data in which the outcome is previously known. The ultimate goal is to identify patterns to help the algorithm predict the outcome when new data is provided. For unsupervised learning on the other hand (B), a random dataset is provided and the ML algorithms try to determine if any existing latent patterns are present.

A.2 Learning Algorithms

A great variety of ML algorithms have been proposed to address specific problems and data formats. These range from more simple implementations, such as Support Vector Machines and Decision Trees, to more complex and robust approaches (Neural Networks), exploring different approaches and capturing different patterns encoded within a dataset.

Support Vector Machines (SVM) algorithm has also proven to be robust and efficient when dealing with classification and regression problems. In classification, its preliminary aim is to establish decision boundaries in the feature space which separate data points belonging to different classes [109] (Figure A.2A). Its intent is to create an optimal separating hyperplane between classes in order to minimize the generalization error and thereby maximise the margin of separating between the classes. When dealing with non-linear separable classes and high number of features, SVM can use non-linear kernel functions to help map high-dimensional feature space. For regression tasks, SVM tries to find the best fit line (hyperplane) that has the maximum number of points within a threshold value (distance between hyperplane and boundary lines).

A Decision Tree is an algorithm that uses a tree-like representation of a dataset in which individual or combination of features are defined as nodes and the possible outcome values from these nodes (branches) are linked to other child nodes, in a process that is repeated until a decision based on a target value can be made. The basic assumption made in the decision tree is that the instances with different classes have different values in at least one of their features. One of the most useful characteristics of such algorithms is their comprehensibility. One can easily understand why the algorithm classifies an instance as belonging to a specific class by just looking at the generated tree and analysing the splitting rules on each node of the tree [107] (Figure A.2B).

Ensemble methods are techniques that combine the results from several machine learning algorithms (base learners) into one predictive model in order to decrease variance, bias and consequently improve predictions. In this regard, Random Forest, one of the most used algorithms in ML uses a combination of decision trees such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the "forest" [110] (Figure A.2D). It is a fast algorithm that

produces highly accurate predictions and can handle a very large number of input variables without overfitting, given that all the trees are built from scratch without any previous information on the other trees in the forest. The final prediction output is given as a combination of the outputs for all the base learners decision trees.

Multi-Layer Perceptron [111], commonly known as MLP, is a feed-forward neural network, consisting of a number of units (neurons), which are connected by weighted links. The units are organized in several layers, namely an input layer, one or more hidden layers, and an output layer. The input layer receives an external activation vector, and forwards it via weighted connections to the units in the first hidden layer. These compute their activations and pass them to the neurons in succeeding layers creating a mapping of the input dataset onto the output vector based on connection weights of the network (Figure A.2D.

Alternatively classification tasks may be addressed with the use of regression algorithms that handle discrete classes (nominal) of the dataset as continuous labels (probability) in a probabilistic classification manner [112]. The classification is then achieved by defining a threshold T, also known as linear decision boundary, in which the predicted classes are assigned. Algorithms that use this type of classification seek for a model that generates the greatest approximate probability function that separates the classes in the dataset [113].

A.3 Feature Selection

When dealing with datasets with a large number of features, precautions are necessary in order to avoid the curse of dimensionality [114] and also model overfitting during the step of training on supervised learning. These two issues introduce a negative effect not only on model generalization but also on performance during training, which could make the predictive task unfeasible [115]. Here, I describe the algorithms used to select the most representative features of a dataset that will be used for this work, before feeding them to the supervised learning algorithms.

Principal Component Analysis (PCA) [116] is a mathematical algorithm that applies an orthogonal transformation in the data, in order to generate a new set of features (principal components) that maximises the variance among the data. Dimensionality

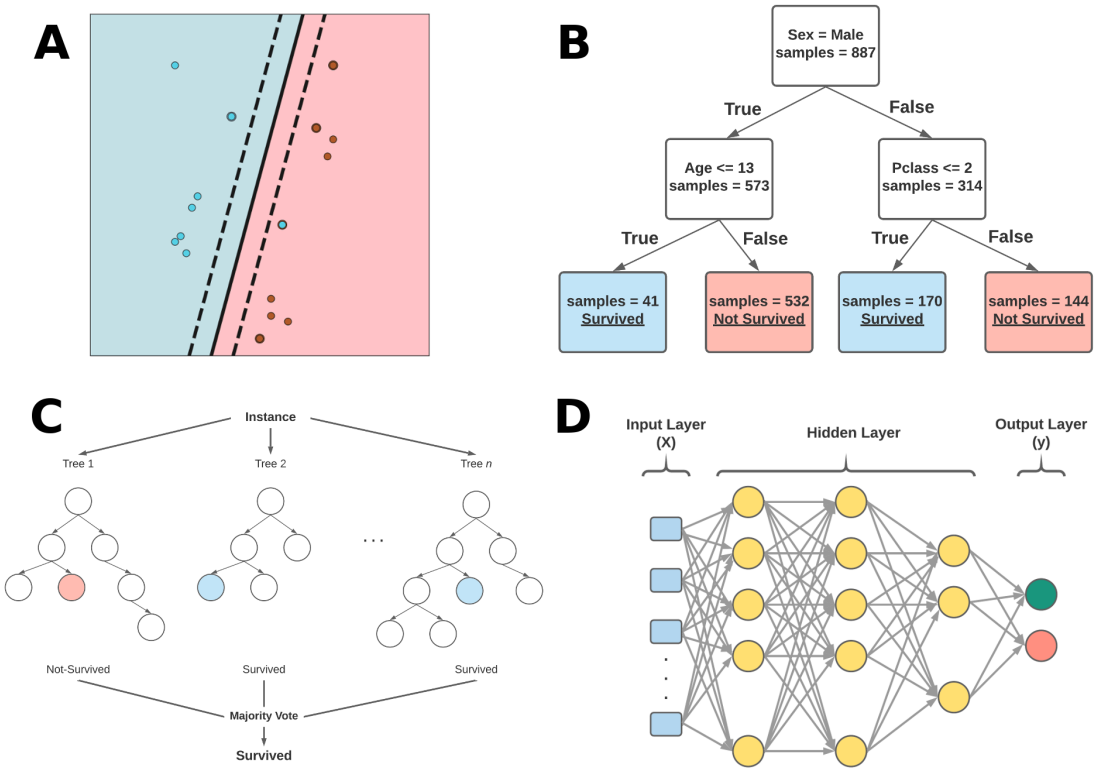


FIGURE A.2: Examples of Learning algorithms workflows. To this date, a variety of ML algorithms have been proposed. More simplistic approaches such as SVM (A) and Decision Tree (B), in particular the latter, are beneficial because, in most cases, they provide a more straightforward and clear explanation of how the algorithm uses the features to make the predictions. One of the most popular ML algorithms is Random Forest, which implements a combination of n base learners decision trees, has proven to be accurate and robust in many different applications. More complex approaches such as Multi-Layer Perceptron are also available, however, in most cases it requires a lot more data and computational power to converge to a non-overfitted solution, than other algorithms.

reduction can then be achieved by selecting only a subset of components with maximal variance (principal components). The latter are also the ones that minimize the mean squared error [117].

Information gain is a technique which uses the assumption that attributes in a dataset are independent and estimates the quality of each attribute by measuring how much information was available before adding that attribute value, and how much was available after with respect to the class that is trying to be predicted [118, 119], commonly calculated in terms of information entropy (Equation A.1). The attributes are then ranked and feature selection can be performed by choosing a subset of features based on their position in the rank.

$$IG(T, X) = Entropy(T) - Entropy(T, X) \tag{A.1}$$

where $IG(T, X)$ represents the information gain towards the target T when adding the attribute X.

More recently, Recursive Feature Elimination (RFE) has been proposed and it uses an external estimator that assigns weights to features (e.g., coefficients of a linear model) to select features by recursively considering smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained. Then, the least important features are pruned from the current set of features [120]. This procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. Alternatively, cross-validation may be performed on each iteration of RFE to find the optimal number of features, also known as RFECV.

A.4 Validation

Predictive models built using ML algorithms aim to generalise from the training data to any data from the problem domain which the model has never "seen" before [107]. This aspect of ML is directly related to the concepts of underfitting and overfitting, which can cause a negative effect of a final predictive model. Underfitting refers to a predictive model unable to identify patterns in the training data and is consequently incapable of generalisation to new data [121]. The solution generally involves trying different algorithms, collection of more data points or generating/modelling of new sets of features. On the other hand, overfitting occurs when an algorithm models the training data so well and is unable to generalise to new data.

The process of evaluating predictive models is crucial in order to obtain a model capable of generalisation and also to estimate its future prediction performance. A commonly used validation method for ML models is known as k-fold cross-validation (CV) [115, 122]. The workflow for this approach consists of randomly splitting a data set (D) into k mutually exclusive subsets, known as folds, $D_1, D_2, ..., D_k$ of approximately equal size. Secondly, the learning algorithm is trained and tested k times; each time $tE1, 2, ..., k$, it

is trained on $D - D_t$ and tested on D_t. The overall performance of the predictive model is then calculated based on the average of performances for each fold (Figure A.3).

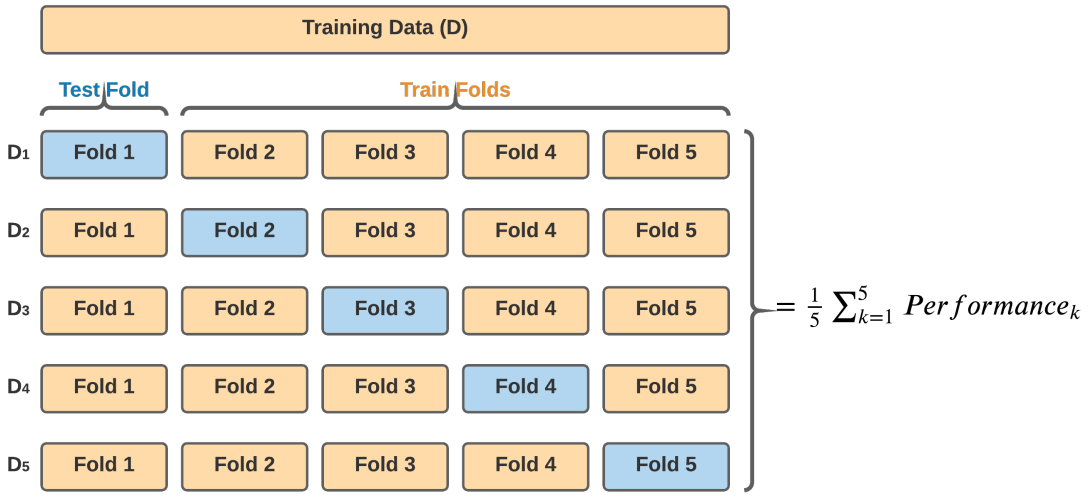


FIGURE A.3: k-fold Cross-validation workflow. The process consists of randomly splitting a data set (D) into k mutually exclusive subsets, known as folds. The learning algorithm is then trained and tested k times and overall performance is calculated based on the average of performances for each fold.

In the case of a dataset with a small amount of data points, the value of k is set to the total number of instances in the dataset, named leave-one-out or jackknife CV [122]. Here the testing set will comprise a single entry and the remaining data is used for training purposes the algorithms. The process is repeated until all entries have been used as a test set.

In addition, as an extra layer of validation for the predictive models independent blind tests are used [51, 55, 73], allowing for the assessment of how likely the final model is to generalise on new data. The proportion of data points on training and blind test sets is directly related to the total size of the dataset.

A.5 Evaluation Metrics

When evaluating performance on classification tasks, a variety of metrics can be expressed based on the values of a contingency table, also known as confusion matrix. Given a binary classification, in which the classes are represented with positive (+) and negative (−), a 2×2 matrix (actual versus predicted class) uses the raw counts of the number of times each predicted label is associated with each real class. As shown on

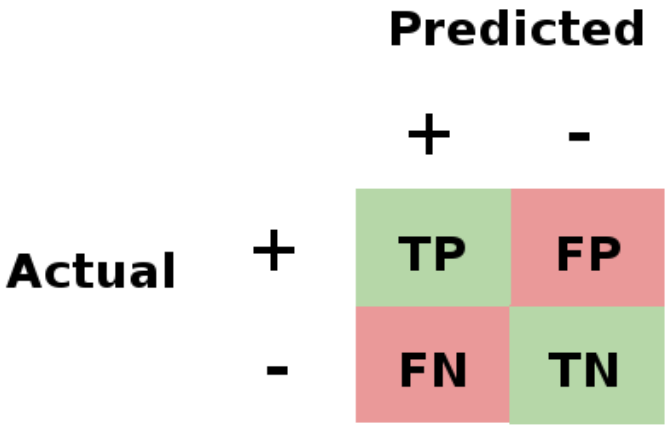


FIGURE A.4: Confusion matrix. True and False Positives (TP and FP) indicate the number of predicted positives that were correct and incorrect, respectively. Similarly, True and False Negatives (TN and FN) refer to correct and wrong predictions for negative class. The sum $TP+FP+TN+FN$ is equal to the total amount number of instances in the data set being used.

Figure A.4, True and False Positives (TP and FP) indicate the number of predicted positives that were correct and incorrect, respectively. Similarly, True and False Negatives (TN and FN) refer to correct and wrong predictions for the negative class.

Alternative metrics can be derived based on the values from the confusion matrix. Precision denotes the proportion of Predicted Positive cases that are Actual Positives (Equation A.2). On the other hand, Recall (Equation A.3) measures the proportion of Actual Positives cases that are correctly Predicted as Positives [123]. Furthermore, a combination of Precision and Recall in a harmonic mean denotes another metric known as F-measure or F-score (Equation A.4) [124]. Precision, Recall and F-score values range from 0 to +1, the latter representing a perfect prediction for the class being evaluated.

By essentially evaluating only one of the classes, all of these metrics present biases towards the predictions of the positive class and ignore the performance in correctly predicting the negative class [124]. This problem can be overcome by a straightforward approach, which consists of inverting the signs on the confusion matrix and recalculating all metrics. In that case, the metrics can be evaluated for each class separately, allowing a better interpretation of the overall predictions. In addition, one can explore a less biased metric, such as the Area Under the ROC Curve (AUC) that considers the True Positive Rate (TPR), also known as sensitivity, which based on the example described previously, corresponds to the proportion of positive data points correctly classified; and also the False Positive Rate (FPR) that corresponds to the proportion of negative data that are wrongly considered as positive, regarding all negative data points. A Receiver Operating Curve (ROC) is then plotted using TPR versus FPR and the AUC is the area under such curve [115].

$$Precision = \frac{TP}{TP + FP} \tag{A.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{A.3}$$

$$Fmeasure = 2 \times \frac{precision \times recall}{precision + recall} \tag{A.4}$$

In the context of regression tasks, Pearson Coefficient Correlation (ρ), Matthews Correlation Coefficient (MCC) [125] and Root Mean Squared Error ($RMSE$) are three well-established and widely used metrics.

ρ is a measure of the linear correlation between two variables X and Y. The coefficient is measured on a scale without units and can take values from -1, total negative linear correlation, to $+1$, total positive correlation, Values closer to 0 indicate that there is no linear correlation between X and Y [126]. The mathematical definition of the Pearson's Correlation Coefficient is given by the covariance of the two variables divided by the product of their standard deviations as described by Equation A.5.

$$\rho_{x,y} = \frac{cov(X, Y)}{\sigma_X \times \sigma_Y} \tag{A.5}$$

where:

- $cov(X, Y)$ is the covariance of X and Y
- σ_X is the standard deviation of the variable X
- σ_Y is the standard deviation of the variable Y

Similarly, MCC also measures the correlation between two variables and also uses the same scale of ρ. MCC scores can be calculated directly from the confusion matrix using the following formula (Equation A.6):

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP) \times (TP + FN) \times (TN + FP) + (TN + FN)} \tag{A.6}$$

$RMSE$ is the standard deviation of the predictions errors. This measure indicates how concentrated the predicted data points are from the line of best fit which represents the ideal perfect correlation between the actual observed values (Y) and the predicted values ($\hat{Y}$) [127]. The mathematical description for this metric is depicted on the Equation A.7.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{A.7}$$

where: n is the total number of instances; $(Y_i - \hat{Y}_i)^2$ represents the squared errors between actual observed values (Y) and the predictions ($\hat{Y}$);

Lastly, the quality of predictions for machine learning algorithms is directly related to the distribution of the features on the dataset being analysed. In this regard, the existence of data points that lie an abnormal distance from other instances of the data, also known as outliers, can decrease the performance of predictive models. These outliers may indicate a variability in the measurements or simply experimental errors. A common practice for predictive methods is to report the performance of their models after removing the outliers and investigate those separately, in order to identify how these abnormal data points may be affecting the final predictions [55, 65].

Appendix B

Programming and Scripting Tools

Task	Tool	Description
Scripting	Python	A high-level programming and scripting language used for general purpose programming. The main Python components used include.
Scripting	R	Programming Language and environment for statistical analysis.
Version Control	Git	A distributed version control system for software development. Used to store and organise code via Bitbucket web service.
Data Manipulation	Numpy	Provides support for large, multi-dimensional arrays, matrices and high-level mathematical functions in Python.
Data Analysis	Pandas	Implementation of data structures and data analysis tools using Python.
Data Analysis	BioPython	Collection of tools for computational biology and bioinformatics in Python.
Machine Learning	Scikit-Learn	Collection of functions and tools to perform Machine Learning in Python.
Sequence clustering and comparison	CD-hit	A program for clustering and comparison of large sets of next generation sequencing data.
Data Storage	SQL	Structured Query Language is a standard language for storing, manipulating and retrieving data in relational databases

Web Framework	Flask	Python framework used for the development of web servers.
Web Development	HTML	Hypertext Markup Language is the standard markup language for creating web pages.
Web Development	CSS	Cascading Style Sheets used for designing an HTML page.
Web Development	JS	JavaScript is used for controlling the behaviour of elements on an HTML page.
Containers	Anaconda	A software distribution with a vast library of pre-configured and pre-built Python packages for scientific computing.
Containers	Docker	Docker provides a way to run applications securely isolated in a container, packaged with all its dependencies and libraries.
Parallelisation	GNU Parallel	Used to run multiple serial computing jobs, such as PDB calculations, in parallel across multiple computing cores, enabling large scale data analysis in reasonable time frames.
High Performance Computing	Spartan	Cluster of computers that delivers high performance. Spartan is a system operated by Research Platform Services (ResPlat) at The University of Melbourne.
High Performance Computing	Bio21	Cluster of computers that delivers high performance. This system is operated by the Bio21 computer staff at The University of Melbourne.

Table B.1: Programing and Scripting tools.

Appendix C

HGDiscovery: an online tool providing functional and phenotypic information on novel variants of homogentisate 1,2-dioxigenase

HGDiscovery: an online tool providing functional and phenotypic information on novel variants of homogentisate 1,2- dioxigenase

Malancha Karmakar[1,2,3,#], Vittoria Cicaloni[1,2,4,#], Carlos H.M. Rodrigues[1,2,3], Ottavia Spiga[4], Annalisa Santucci[4], David B. Ascher[1,2,4,5,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[2] Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne, Melbourne, Victoria, Australia

[3] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia

[4] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Siena, Italy

[5] Department of Biochemistry, Bio21 Institute, University of Cambridge, Cambridge, UK

# These authors contributed equally.

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au.

Abstract

Alkaptonuria (AKU), a rare genetic disorder, is characterized by the accumulation of homogentisic acid (HGA) in the body. Affected individuals lack enough functional levels of an enzyme required to breakdown HGA. Mutations in the *HGD* gene cause AKU and they are responsible for deficient levels of functional homogentisate 1,2-dioxygenase (HGD), which, in turn, leads to excess levels of HGA. Although HGA is rapidly cleared from the body by the kidneys, in the long term it starts accumulating in various tissues, especially cartilage. Over time (rarely before adulthood), it eventually changes the color of affected tissue to slate blue or black. Here we report a comprehensive mutation analysis of 111 pathogenic and 190 non-pathogenic HGD missense mutations using protein structural information. Using our comprehensive suite of graph-based signature methods, mCSM complemented with sequence-based tools, we studied the functional and molecular consequences of each mutation on protein stability, interaction and evolutionary conservation. The scores generated from the structure and sequence-based tools were used to train a supervised machine learning algorithm with 84% accuracy. The empirical classifier was used to generate the variant phenotype for novel HGD missense mutations. All this information is deployed as a user friendly freely available web server called HGDiscovery (http://biosig.unimelb.edu.au/hgdiscovery/).

Introduction

Alkaptonuria (AKU) is a rare recessive metabolic disorder which was used by Sir Archibald Garrod in his Croonian lectures to describe inborn errors of metabolism [1]. It is a hereditary disorder, resulting from mutations in the enzyme homogentisate 1,2 dioxygenase (HGD) (EC 1.13.11.5), responsible for the breakdown of homogentisic acid (HGA) which is an intermediate metabolite in the tyrosine degradation pathway [2]. With blockage in tyrosine metabolism, elevated levels of HGA leads to deposition of its own polymers as an ochronotic pigment in the connective tissue including cartilage, heart valves, and sclera [3]. Manifestation of disease during early childhood is seen as "homogentisic aciduria", which is darkening of the urine upon standing. Delayed symptoms can be seen after 30 years of age which involves "ochronosis" – pigmentation of collagenous tissues like cardiac valves, eyes, ears and skin [4]. Current estimates of the disease occurrence in the Unites States obtained from the National Organisation of Rare Disorders is 1 in 250,000 – 1,00,000 live births [5].

HGD gene located on chromosome 3q21-q23 [6], is a single copy gene composed of 14 exons [7]. Due to compound heterozygosity or homozygosity of HGD gene variants, the enzymatic defect in HGD is autosomal recessive [6, 8]. Information on all variants identified till date globally have been documented in the HGD mutation database (http://hgddatabase.cvtisr.sk/). The experimental crystal structure of the HGD protein has been solved (PDB code 1EY2 and 1EYB) in 2000. The HGD protein protomer (NP_000178.2), is composed of 445 amino acids, which includes a 280 residue N-terminal domain, a central β-sandwich and a 140 residue C-

terminal domain [8]. It is a complex hexameric protein arranged as a dimer of trimers [9]. It is principally expressed in osteoarticular compartment cells (i.e. chondrocytes, synoviocytes and osteoblasts) [10] and in prostate, small intestine, colon, kidney and liver [7]. The spatial structure of the protomer, two-disc like trimers and the hexamer are maintained by an intricate network of non-covalent inter and intra-molecular interaction. This makes the protein structure extremely vulnerable to mutations [11].

The major obstacle in studying an ultra-rare and complex disease like AKU is the lack of a standardized methodology to assess disease severity and response to treatment [12], which is complicated by the fact that AKU symptoms differ from one individual to another. Detailed evaluation and comparison of clinical and genomic data of AKU patient can play a key role to understand AKU variability. An in-depth molecular characterization of the disease is needed in pharmacogenomics prediction for suitable medical treatment. To address the issue we developed ApreciseKUre platform, which includes data on potential biomarkers, patients' quality of life, biochemical outcomes and clinical information facilitating their integration and analysis in order to shed light on pathological characterization of every AKU patient in a typical Precision Medicine perspective [13-16] .

We wanted to further elaborate and build a new database which would complement the existing ApreciseKUre database. The new database would provide the necessary underlying molecular information for novel and known clinical HGD variants. We have tried to exploit structural and sequence based information to build a predictive tool using supervised machine

learning algorithm. The model has been implemented through the webserver HGDiscovery, providing functional and phenotypic consequences of HGD non-synonymous variations to better guide clinical decisions.

Methods

Data curation

After removal of duplicate mutations, we curated a dataset composed of 301 non-synonymous substitutions. It included 190 non-pathogenic non-synonymous variations retrieved from gnomAD v.3 (Genome build GRCh38/hg38, Ensembl gene ID: ENSG00000113924.11, Region 3:120628173-120682571) [17] and 111 AKU-causing clinical mutations. The 111 variants were first described in the study of Ascher et al. 2019 [18] and included in HGD Mutation Database (http://hgddatabase.cvtisr.sk) [19], which summarizes results of mutation analysis from approximately 530 AKU patients reported so far.

HGD protein structure

The X-ray crystallographic 3D structure of *Homo sapiens* holo-HGD (holo-HGDHs, PDB ID: 1EY2) is incomplete; thus, it needed structural reconstruction of the missing residues of the monomer and then of the whole hexamer in order to be able to perform a complete evaluation of variants effect on protein stability and flexibility. The missing loop in the human protein structure (residues 348–355) was reconstructed by homology modeling using the *Pseudomonas putida* HGD (HGDPp) structure. By using protein BLAST [20] software we found three structures belonging to *Pseudomonas putida* with a sequence identity (the amount of characters which

match exactly between two different sequences) larger than 49% and with root-mean-square deviation (RMSD) amounting to 1.8 Å for Cα [21]. We opted for HGDPp, with PDB ID 4AQ2 since, similarly to 1EY2, as it had no substrate. The structures of holo-HGDHs (PDB ID: 1EY2) and its homologous HGDPp (PDB ID: 4AQ2) were retrieved from the Protein Data Bank (PDB) [22]. Thereafter at the 1EY2 and 4AQ2 sequences alignment on BLAST web server [20], we modelled the missing residues. The modelling of the loop 348-355 was carried out using a homology model approach in which an elucidated structure of HGDPp loop was employed as template to model the structure of the protein of interest. The completed monomer structure served as a starting point for the reconstruction of the whole HGDHs oligomeric protein on the template of the asymmetric units of PDB entry 1EY2. The structure reliability was validated using PROCHECK [23]. Additionally, the energy minimization of the hexameric protein was performed using GROMACS 5.0.2 [24] in order to obtain an optimized 3D structure, a relaxation of the highly energetic conformations and a correct geometry for the following simulations (for additional information see Supplementary Methods in [18]).

Biophysical and evolutionary score generation

A thorough structural and sequence based assessment was performed for all the HGD variants to account for the potential effects of AKU-causing mutations. Variations in protein-protein interactions between the different monomers of the hexamer HGD upon mutation was determined using mCSM-PPI2 [25]. Changes in protein stability and folding were determined using our in-house tools like mCSM-Stability [26], SDM [27] and DUET [28] and conformational flexibility changes using the normal mode analysis tool called DynaMut [29]. Effects of

mutations on binding affinity of HGD to its substrate homogentisic acid were analyzed using mCSM-Lig [30]. All these are novel machine learning approaches that use graph-based signatures to represent the structural and biochemical environment of the wild-type 3D structure of a protein to quantitatively predict the effects of point mutation. To complement the above methods we used sequence based feature like SNAP2 (Screening for Non-Acceptable Polymorphisms) [31], ConSurf [32] and Provean (Protein Variation Effect Analyzer) [33] which provides valuable evolutionary information. To enrich the analysis we included protein's wild type structural information such as residue depth, dihedral angles of the HGD chain ϕ (phi) and ψ (psi), relative solvent accessibility and secondary structure information. We calculated changes in molecular interactions such as hydrophobic, ionic, van der Waals', halogen and hydrogen bonds and π interactions (cation–π, donor–π, halogen–π, carbon–π, π–π) between the wild type and mutant structures using Arpeggio [34]. We also included population-based variability using the missense tolerance ratio (MTR) [35] scoring system.

Supervised Machine learning for empirical model building

We evaluated different supervised machine learning algorithms for classification which is available within the scikit-learn Python library. These include – K-Nearest Neighbors (KNN), Random Forest, Decision Trees, Extra Trees, AdaBoost, Gradient Boosting, SVM, Gaussian Naïve Bayes, and Stochastic Gradient Descent. The best performing model was chosen by assessing metrics like Matthews correlation co-efficient (MCC), Receiver Operating Characteristic (AUROC) curve, accuracy, F1-score and precision. The model was trained using stratified 10-fold

cross validation. We carefully split the train and blind test dataset non-redundantly with respect to the amino acid residue position.

To address the issue of imbalance between the pathogenic and non-pathogenic mutations in the data, we evaluated the model performance by both under-sampling the non-pathogenic mutations and oversampling pathogenic mutations in the train dataset [36]. The performance was compared for above mentioned scenario and the normal dataset and best results were obtained when the pathogenic mutations were oversampled using the Extra Tree algorithm. **Ex**tremely **ra**ndomized **tree** classifier (or Extra Tree) is an ensemble machine learning algorithm and a variation of the random forest algorithm. The empirical binary classifier built using this algorithm highlights a set of structural and evolutionary features which can be used to discriminate between AKU-causing and non-pathogenic variations.

Webserver development

We have implemented HGDiscovery as a user-friendly and freely available webserver (http://biosig.unimelb.edu.au/hgdiscovery/). The front-end of the server was developed using Materializecss framework version 1.0.0, while the back-end was built in Python using the Flask framework version 1.0.2. The server is hosted on a Linux server running Apache 2.

Results

In this work we have used the 3D protein structure to understand the functional and molecular consequences of mutations in HGD leading to AKU disease and using the information generated

from these analyses we have trained a supervised machine learning algorithm to develop a predictive tool to determine novel variants which could lead to AKU manifestation. Figure 1 depicts the novel methodological pipeline we have developed.

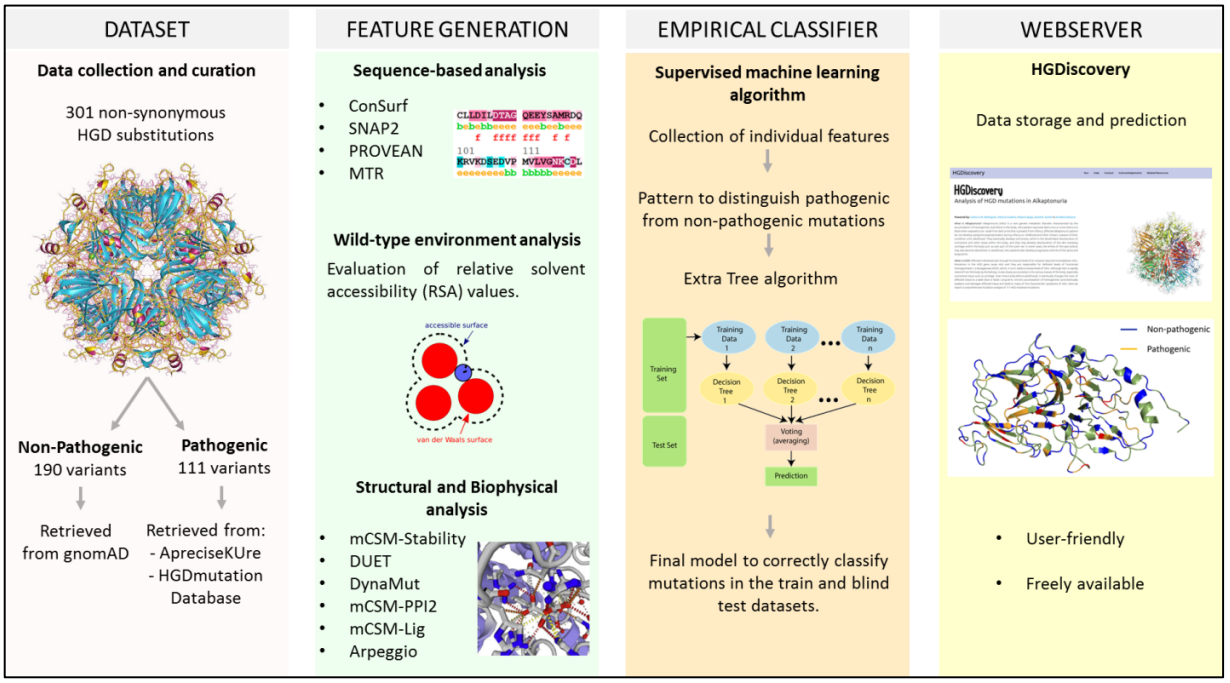


Figure 1: HGDiscovery workflow. *The first step involves scoping published literature and clinical databases to prepare a curated list of non-synonymous HGD mutations. The second step involves generating various structure and sequence-based features for the curated missense mutations. In the third step, we use these features in a supervised machine learning algorithm to build a binary classifier, which can distinguish between pathogenic and non-pathogenic missense mutations. Finally, we develop a free available user-friendly webserver which contains phenotypic information on all HGD variants.*

Sequence-based analysis of HGD variants

ConSurf, SNAP2 and PROVEAN are sequence-based predictors and consider evolutionary information to predict functionally important non-synonymous mutation. The prediction helps us understand the biological impact of a mutation on the protein structure. A consistent pattern was observed from all of the sequence based features. The pathogenic mutations were associated with deleterious scores and the non-pathogenic mutations scored neutral. All the features were statistically significant to be used to train the predictive algorithm to build the empirical tool (p-values SNAP2: 4.6 e^{-14}, PROVEAN: 1.1 e^{-9}, ConSurf: 2.4 e^{-10}). Population-based variability was considered using the missense tolerance ratio (MTR) scoring system. Majority of the pathogenic mutations were in the bottom 25th percentile, reflecting intolerance and hence associated with altering protein function.

Wild-type environment analysis

The wild-type environment analysis includes data on relative solvent accessibility (RSA), residue depth, dihedral angles and secondary structure information for both pathogenic and non-pathogenic variants. Looking into the relative solvent accessibility values for the pathogenic and non-pathogenic mutations (p-value: 2.2 e^{-8}), we see pathogenic mutations tend to be more exposed than non-pathogenic variants. It has been previously described that the HGD protomer structure constitutes of a pore in which the side chains of large number of residues are exposed [21]. These residues are thought to play an important part in the complex HGD catalytic function and we see subtle changes in the side chains as non-synonymous substitution can affect the active site functionality [18]. The residue depth values reveal pathogenic mutations are more buried than non-pathogenic mutations. This observation is congruous with earlier

observation where point mutations on the surface were better tolerated in the globular hexameric HGD protein structure.

Structural and Biophysical analysis

Our in-house biophysical tools mCSM-Stability [26], DUET [28] and DynaMut [29] were used to study and understand the impact of missense mutations on protein stability, folding and conformational flexibility. These tools are novel machine-learning algorithms which rely on graph-based signatures to calculate changes in Gibb's free energy upon non-synonymous mutations. We observed pathogenic mutations to be associated with highly destabilizing scores affecting protein stability and dynamics. The effects of mutation on the substrate binding affinity to active site were determined using mCSM-Lig [30]. Pathogenic mutations altered the active / substrate binding pocket. mCSM-PPI2 [25] was used to assess changes in protein-protein interaction and we observed pathogenic mutations hindered the formation of the symmetrical homohexamer. Therefore, pathogenic mutations either reduced or disrupted the HGD protein activity.

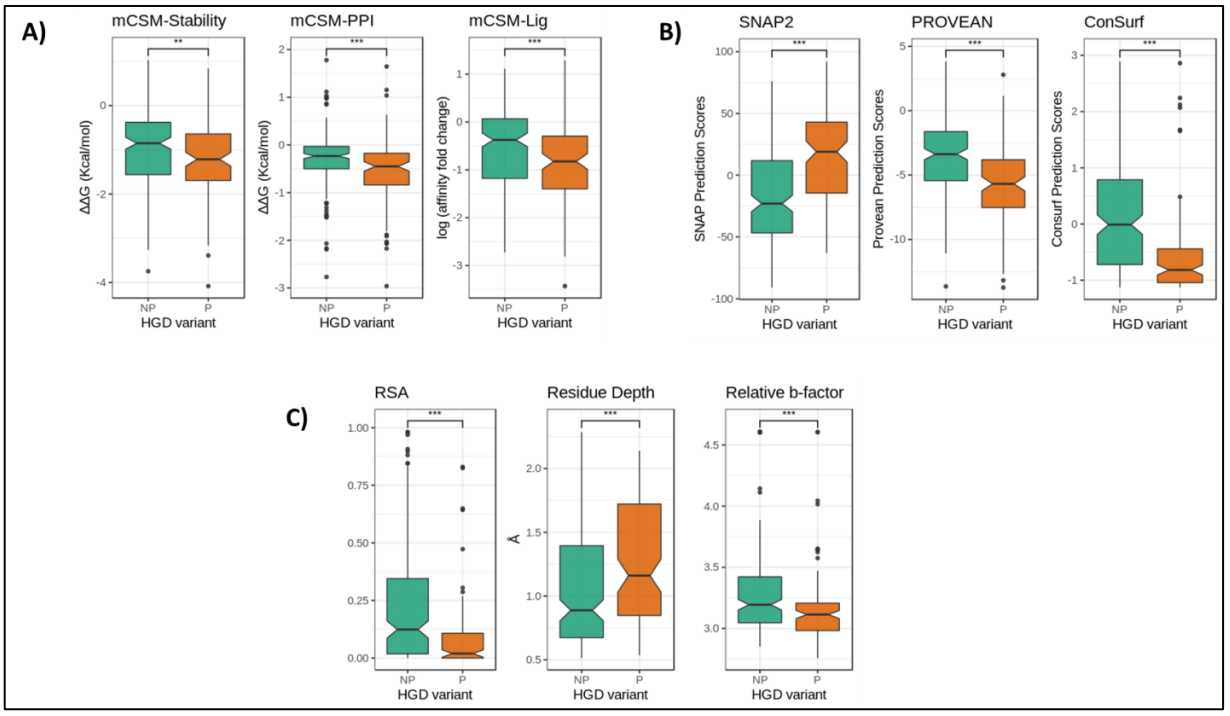


Figure 2: Boxplot representation of features. *A) Structural features. B) Sequence-based features. C) Wild-type environment features. The non-pathogenic mutations (NP) are represented as sea green and pathogenic mutations (P) as dark orange. (*** p < 0.0001, ** p < 0.001, Welch two sample t-test).*

Supervised machine learning algorithm: Extra Tree

Our features could be grouped into eight distinct categories – protein stability, protein-protein interactions, ligand affinity, evolutionary conservation scores, distance parameters, MTR scores, molecular interaction and backbone geometry. Each category of features was initially used to build and evaluate the performance of the predictive model. After a thorough analysis of the individual features, we combined them together to see if there is a pattern which could be used to distinguish pathogenic from non-pathogenic HGD mutations. We observed that when

different categories of features were combined together, in addition to using stratified 10-fold cross validation with Extra Tree algorithm, yielded a more robust and balanced performance. The Extra Tree algorithm implements a meta estimator that fits randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and reduces over-fitting [37].

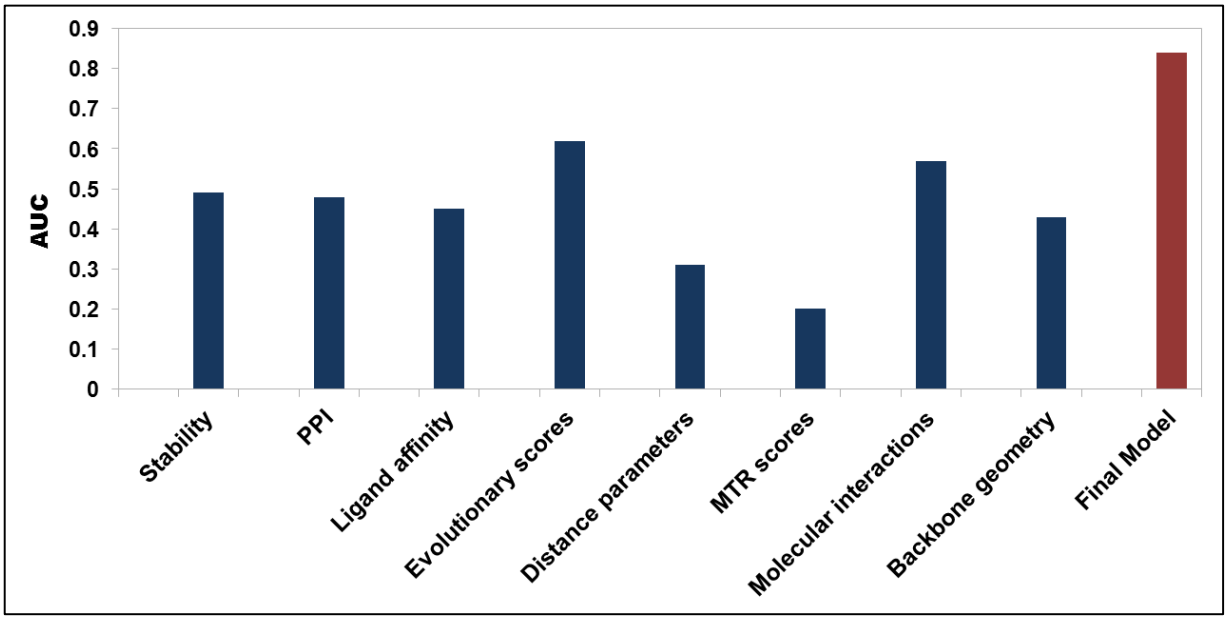


Figure 3: Empirical model performance trained on individual class of features. *The Extra Tree algorithm was trained using stratified 10-fold cross validation using eight distinct class of features (first eight bars from left to right; dark blue bars) and with a combination of all features (red bar). The AUC score is low when a single class of feature is used for training the binary classifier, however, a significant improvement is noticed when all the eight different features are combined to build the model.*

190 non-pathogenic and 111 pathogenic mutations were split into non-redundant train and blind test datasets with respect to their amino acid position. Initially we observed poor performance on the model's ability to predict pathogenic mutation. We concluded that the train data set was imbalanced as there were more non-pathogenic mutations than pathogenic mutations. We improved the metric scores by oversampling (duplicating) [36] the pathogenic mutations in the train dataset. The final model correctly classified 84% and 73% of mutations in the train and blind test datasets respectively.

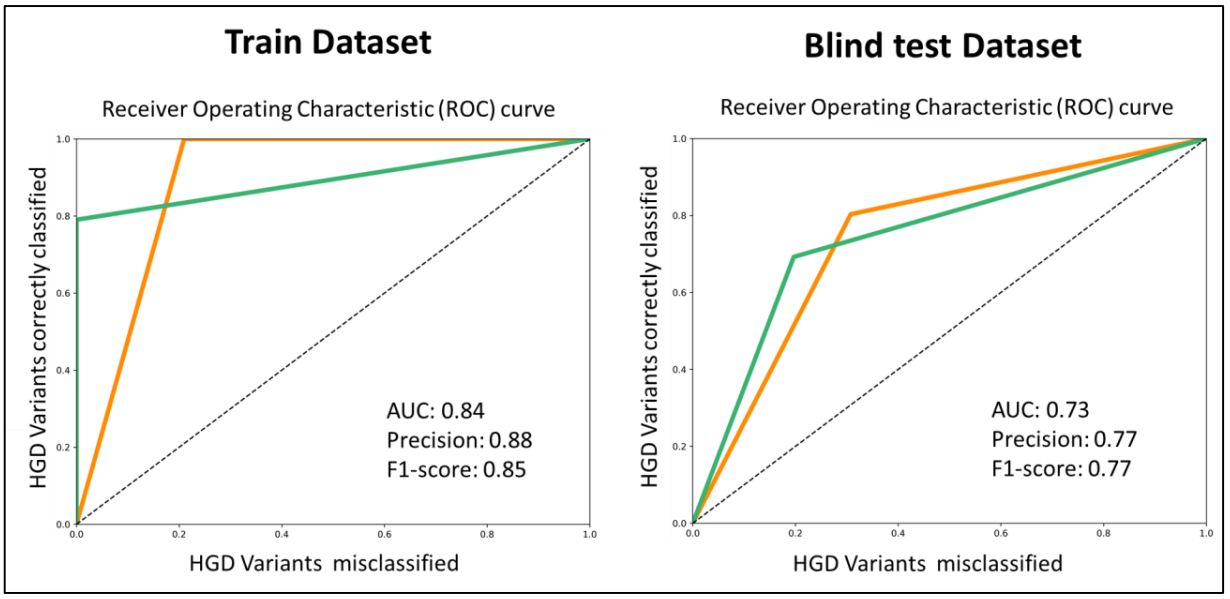


Figure 4: Receiver Operating Characteristic (ROC) curves of HGD classifier. *The evaluation metrics shown for train and test dataset where pathogenic mutations are represented in dark orange and non-pathogenic mutations in sea green. (AUC = area under the curve).*

HGDiscovery Webserver

HGDiscovery allows for users to query for a single point mutation or submit a list of mutations to be analysed in batch. For the "Single Mutation" option users are asked to provide the point

mutation as a string containing the wild-type reside one-letter code, its corresponding residue number and the mutant residue one-letter code. The "Mutation List" option requires that a text file is submitted with the list of mutations (one per line).
The results page for the "Single Mutation" option displays the predicted outcome on the top alongside with details of the input mutation, wild-type residue environment, the variables and scores used by our predictive model and external links to experimental evidence (when available). An interactive 3D viewer using the NGL-viewer [38] shows the molecular contacts generated by Arpeggio [34] for wild-type and mutant structures.

On the "Mutation List" option, the results are displayed as a downloadable table. Individual analysis for each variant on the table can be analysed similarly to "Single Mutation" option by clicking the "Details" button. An interactive viewer is also shown at the bottom of the page highlighting Pathogenic and Non-pathogenic mutations on the 3D structure.

Discussion

Here we present an empirical classifier HGDiscovery, which has phenotypic information on all variants of homogentisate 1,2 dioxygenase, (EC 1.13.11.5), an enzyme involved in the metabolism of tyrosine, whose deficiency leads to Alkaptonuria [OMIM 203500]. We combine structural, evolutionary and molecular information from known HGD variations and look to investigate a pattern to distinguish non-pathogenic from AKU-causing non-synonymous variants. So along with physiological information from ApreciseKUre platform, we have an additional AKU-dedicated database which provides new insight into functional and phenotypic

consequences of novel HGD non-synonymous variations, crucial for a genetic disease like AKU to support clinical decisions.

The 3D crystal structure of the HGD active form reveals a highly complex and dynamic hexameric organization comprising two disk-like trimers [9]. An intricate network of noncovalent interactions is needed to maintain the spatial structure firstly of the protomer, the trimer and then the hexamer. This delicate structure presents a very low tolerance to mutations and can be easily disrupted mainly by missense variants compromising enzyme function. In case of HGD, missense variants represent approximately 65% of all known AKU substitutions [4, 11, 39] and 93 distinct amino acid residue positions within the structure are affected by the 111 AKU-causing missense changes. Recent studies on evolutionary conservation revealed that AKU variants were mainly located at more conserved residue positions [18] and, consequently, HGD missense changes can influence protein folding and stability or interactions with other protomers or substrate. Specifically, they can decrease stability of individual protomers, disrupt protomer–protomer interactions, or modify residues in the active-site region. Thus, when a novel HGD missense mutation is identified, it is important to distinguish causal AKU variants from non-pathogenic ones.

With sequence-based tools such as ConSurf, SNAP2 and PROVEAN we have evaluated evolutionary information in order to predict functionally important non-synonymous mutations and the biological impact of a mutation on HGD protein structure. The obtained results supported our hypothesis: the pathogenic mutations were associated with deleterious scores whereas the non-pathogenic mutations with neutral scores. Additionally, using MTR score

system we have analyzed population-based variability and most of the pathogenic mutations resulted to be in the bottom 25th percentile, reflecting intolerance and alteration of protein function. With the help of biophysical tools (i.e. mCSM-Stability, DUET and DynaMut) we investigated the impact of missense mutations on protein stability, folding and conformational flexibility. AKU-causing mutations appear to reduce or disrupt the HGD protein activity by destabilizing its structure and altering the active site/substrate binding pocket.

It is not uncommon that AKU patients carry compound heterozygotes for two HGD gene variants. In such cases, the estimation of the role of each missense variant is not trivial, since the hexamer could be assembled with monomers all affected by the same variant (homo-oligomer) or by two different ones (heterooligomer) [40]. Variants affecting two different regions could have additive destructive effect, on the contrary, the effects could partially compensate for those that belong to the same region. However, we do not have any tools able to evaluate such events so far [12]. Compound heterozygosity could have even interfered with our analysis, where a variant labelled as non-pathogenic could actually be pathogenic. This was the limitation of our study. But with increasing availability of genomic and clinical data after patient analysis in future, we can always update our tool and re-label the mislabeled non-synonymous variants.

The information available from the above study can be used to develop new treatment strategies, for example, use of small molecules. We know that a pathogenic mutation with destabilizing scores for stability and flexibility leading to reduced enzyme activity can be rescued partially or totally with the help of a small molecule and hence might decrease the

severity of the disease [18]. Moreover, understanding the protein structure and function would also help in designing tailored drugs and therapies.

Therefore, this framework may represent an online tool that can be turned into a best practice model for Rare Diseases. We believe this is not limited to the study of AKU, but it represents a proof of principle study that could be applied to other rare diseases, allowing data management, analysis and interpretation. We applied this novel methodological pipeline to understand and determine novel drug resistant mutations in tuberculosis [41, 42] and even performed a real-time analysis [43] on tuberculosis patient. Hence, HGDiscovery is a user friendly freely available tool which could help with faster and more accurate diagnosis of AKU.

Acknowledgements

M.K and C.H.M.R were funded by Melbourne Research Scholarships. D.B.A. was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council and Fundacao de Amparo a Pesquisa do Estado de Minas Gerais(FAPEMIG) [MR/M026302/1]; the Jack Brockhoff Founda-tion [JBF 4186, 2016]; and an Investigator Grant from the National Health and Medical Research Council of Australia[GNT1174405]. Supported in part by the Victorian Government's OIS Program.

References

1. Garrod, A.E., *The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article].* The Yale journal of biology and medicine, 2002. **75**(4): p. 221-231.
2. Phornphutkul, C., et al., *Natural history of alkaptonuria.* N Engl J Med, 2002. **347**(26): p. 2111-21.
3. Damarla, N., et al., *Alkaptonuria: A case report.* Indian journal of ophthalmology, 2017. **65**(6): p. 518-521.
4. Zatkova, A., L. Ranganath, and L. Kadasi, *Alkaptonuria: Current Perspectives.* Appl Clin Genet, 2020. **13**: p. 37-47.
5. Disorders, N.O.f.R., NORD, 2019. **https://rarediseases.org/rare-diseases/alkaptonuria/**.
6. Pollak, M.R., et al., *Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2.* Nat Genet, 1993. **5**(2): p. 201-4.
7. Fernandez-Canon, J.M., et al., *The molecular basis of alkaptonuria.* Nat Genet, 1996. **14**(1): p. 19-24.
8. Janocha, S., et al., *The human gene for alkaptonuria (AKU) maps to chromosome 3q.* Genomics, 1994. **19**(1): p. 5-8.
9. Titus, G.P., et al., *Crystal structure of human homogentisate dioxygenase.* Nat Struct Biol, 2000. **7**(7): p. 542-6.
10. Laschi, M., et al., *Homogentisate 1,2 dioxygenase is expressed in human osteoarticular cells: implications in alkaptonuria.* J Cell Physiol, 2012. **227**(9): p. 3254-7.
11. Nemethova, M., et al., *Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy.* Eur J Hum Genet, 2016. **24**(1): p. 66-72.
12. Ranganath, L.R. and T.F. Cox, *Natural history of alkaptonuria revisited: analyses based on scoring systems.* J Inherit Metab Dis, 2011. **34**(6): p. 1141-51.
13. Cicaloni, V., et al., *Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease.* Faseb j, 2019. **33**(11): p. 12696-12703.
14. Spiga, O., et al., *Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease.* Orphanet J Rare Dis, 2020. **15**(1): p. 46.
15. Spiga, O., et al., *A new integrated and interactive tool applicable to inborn errors of metabolism: Application to alkaptonuria.* Comput Biol Med, 2018. **103**: p. 1-7.
16. Spiga, O., et al., *ApreciseKUre: an approach of Precision Medicine in a Rare Disease.* BMC Med Inform Decis Mak, 2017. **17**(1): p. 42.
17. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
18. Ascher, D.B., et al., *Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU.* Eur J Hum Genet, 2019. **27**(6): p. 888-902.
19. Zatkova, A., et al., *Identification of 11 Novel Homogentisate 1,2 Dioxygenase Variants in Alkaptonuria Patients and Establishment of a Novel LOVD-Based HGD Mutation Database.* JIMD Rep, 2012. **4**: p. 55-65.
20. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
21. Jeoung, J.-H., et al., *Visualizing the substrate-, superoxo-, alkylperoxo-, and product-bound states at the nonheme Fe(II) site of homogentisate dioxygenase.* Proceedings of the National Academy of Sciences, 2013. **110**(31): p. 12625.
22. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

23. Laskowski, R., et al., *PROCHECK: A program to check the stereochemical quality of protein structures.* Journal of Applied Crystallography, 1993. **26**: p. 283-291.
24. Berendsen, H.J.C., D. van der Spoel, and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation.* Computer Physics Communications, 1995. **91**(1): p. 43-56.
25. Rodrigues, C.H.M., et al., *mCSM-PPI2: predicting the effects of mutations on protein–protein interactions.* Nucleic Acids Research, 2019. **47**(W1): p. W338-W344.
26. Pires, D.E.V., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics (Oxford, England), 2014. **30**(3): p. 335-342.
27. Pandurangan, A.P., et al., *SDM: a server for predicting effects of mutations on protein stability.* Nucleic Acids Res, 2017. **45**(W1): p. W229-w235.
28. Pires, D.E.V., D.B. Ascher, and T.L. Blundell, *DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach.* Nucleic acids research, 2014. **42**(Web Server issue): p. W314-W319.
29. Rodrigues, C.H.M., D.E.V. Pires, and D.B. Ascher, *DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability.* Nucleic Acids Research, 2018. **46**(W1): p. W350-W355.
30. Pires, D.E., T.L. Blundell, and D.B. Ascher, *mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance.* Sci Rep, 2016. **6**: p. 29575.
31. Hecht, M., Y. Bromberg, and B. Rost, *Better prediction of functional effects for sequence variants.* BMC Genomics, 2015. **16 Suppl 8**: p. S1.
32. Ashkenazy, H., et al., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Res, 2016. **44**(W1): p. W344-50.
33. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015. **31**(16): p. 2745-2747.
34. Jubb, H.C., et al., *Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures.* Journal of molecular biology, 2017. **429**(3): p. 365-371.
35. Traynelis, J., et al., *Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation.* Genome Res, 2017. **27**(10): p. 1715-1729.
36. Krawczyk, B., *Learning from imbalanced data: open challenges and future directions.* Progress in Artificial Intelligence, 2016. **5**(4): p. 221-232.
37. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine Learning, 2006. **63**(1): p. 3-42.
38. Rose, A.S. and P.W. Hildebrand, *NGL Viewer: a web application for molecular visualization.* Nucleic Acids Research, 2015. **43**(W1): p. W576-W579.
39. Zatkova, A., *An update on molecular genetics of Alkaptonuria (AKU).* J Inherit Metab Dis, 2011. **34**(6): p. 1127-36.
40. Gallagher, J.A., et al., *Alkaptonuria: An example of a "fundamental disease"--A rare disease with important lessons for more common disorders.* Semin Cell Dev Biol, 2016. **52**: p. 53-7.
41. Karmakar, M., et al., *Structure guided prediction of Pyrazinamide resistance mutations in pncA.* Scientific Reports, 2020. **10**(1): p. 1875.
42. Karmakar, M., et al., *Empirical ways to identify novel Bedaquiline resistance mutations in AtpE.* PLoS One, 2019. **14**(5): p. e0217169.
43. Karmakar, M., et al., *Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy.* Am J Respir Crit Care Med, 2018. **198**(4): p. 541-544.

Appendix D

MTR3D: Identifying regions within protein tertiary structures under purifying selection

MTR3D: Identifying regions within protein tertiary structures under purifying selection

Michael Silk[1,2,3], Douglas Pires[1,2,3,4], Carlos M. Rodrigues[1,2,3], Elston N. D'Souza[1,2,3], Moshe Olshansky[1], Natalie Thorne[5], David B. Ascher[1,2,3,6,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Australia
[2] Structural Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Melbourne, Australia
[3] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Australia
[4] School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[5] Melbourne Genomics Health Alliance, Melbourne, Australia
[6] Department of Biochemistry, University of Cambridge, Cambridge, UK

[*] To whom correspondence should be addressed D. B. A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au.

Abstract

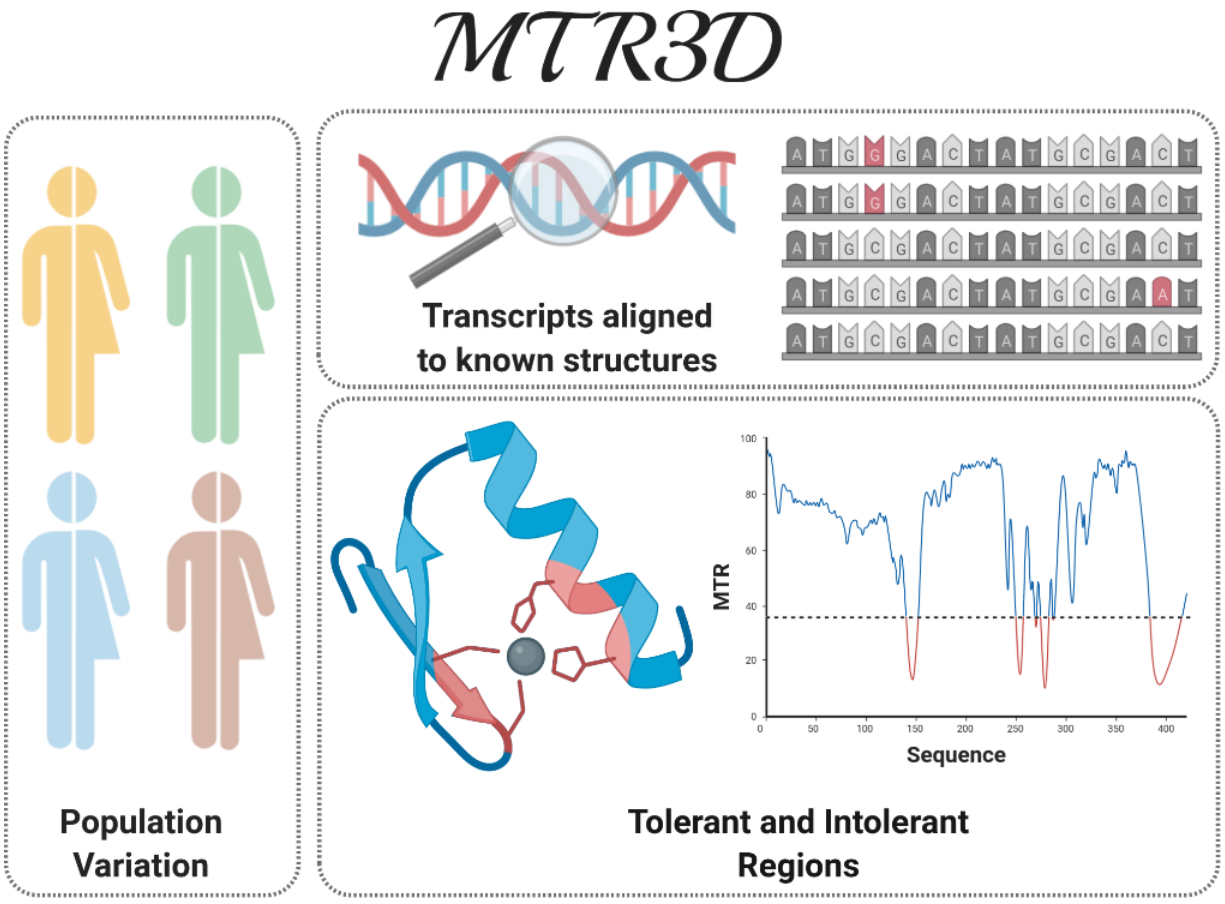
The identification of disease-causal variants is non-trivial. By mapping population variation from over 448,000 exome and genome sequences to over 81,000 experimental structures and homology models of the human proteome, we have calculated both regional intolerance to missense variation (Missense Tolerance Ratio, MTR), using a sliding window of 21-41 codons, and introduce a new 3D spatial intolerance to missense variation score (3D Missense Tolerance Ratio, MTR3D), using spheres of 5-8 Å. We show that the MTR3D is less biased by regions with limited data and more accurately identifies regions under purifying selection than estimates relying on the sequence alone. Intolerant regions were highly enriched for both ClinVar pathogenic and COSMIC somatic missense variants (Mann-Whitney U test $p < 2.2 \times 10^{-16}$). Further, we combine sequence- and spatial-based scores to generate a consensus score, MTRX, which distinguishes pathogenic from benign variants more accurately than either score separately (AUC = 0.85). The MTR3D server enables easy visualisation of population variation, MTR, MTR3D and MTRX scores across the entire gene and protein structure for >17,000 human genes and >42,000 alternative alternate transcripts, including both Ensembl and RefSeq transcripts. MTR3D is freely available by user-friendly web-interface and API at http://biosig.unimelb.edu.au/mtr3d/ .

Keywords: Diagnostic tool, mutational analysis, missense variant intolerance, population variation; in silico score; missense constraint

Key Points:

- MTR3D captures spatial-based intolerance to identify regions mapped to protein 3D structures under purifying selection enriched for disease-causing missense variants and functionally important regions
- A consensus score combines sequence-based and spatial-based intolerance and achieves greater predictive power of disease variants
- The MTR3D web-server provides an easy-to-use tool to view MTR3D scores across experimental and homology modelled protein structures, with the Ensembl and RefSeq transcript sets mapped to the structures.

Graphical Abstract



INTRODUCTION

Advancements in our ability to distinguish between pathogenic and benign variants using both experimental and computational methods have proven greatly beneficial in our ability to diagnose genetic diseases. *In silico* predictors of deleteriousness have been successfully used to prioritise likely disease-causative variants (1-3), and have proven greatly beneficial in a number of disease cohorts, such as epilepsy, to identify genes enriched for pathogenic variation (4). Despite the accuracy of these methods improving, it remains challenging to identify causative variants due to the diverse effects that a mutation can have on the resulting protein.

Large publicly available datasets of observed variation within the population provide the means to identify regions under purifying selection that are intolerant to genetic change. Several methods have been used to measure this using sequence-based approaches, including RVIS (5), MPC (6) and MTR (7), which have shown that patient-ascertained causative variants are preferentially found within intolerant regions. These provide differing scores depending on whether they are per-gene or regional scores, the sample sizes involved, and the statistical methods used to summarise the degree of depletion. Several tools exist that utilise sequence-based information mapped to protein tertiary structures in order to analyse the impact of mutations (8,9). When examining intolerance scores across a gene's protein tertiary structure, intolerant regions have been observed to cluster within 3D space, but this has not been systematically and comprehensively investigated.

To form a more accurate estimate of missense intolerance, and to systematically investigate genetic intolerance in the tertiary protein space, we introduce the MTR3D, a means of evaluating missense variant deleteriousness based on its spatially measured intolerance. The MTR3D provides both experimental structures from the Protein Data Bank (PDB) and available homology models where a transcript (Ensembl or NCBI RefSeq) could be aligned to a high-quality template.

MATERIALS AND METHODS

Data sets

Population variation was combined from gnomAD v2.1.1 (10) (125,748 exomes, 15,708 genomes), gnomAD v3 (76,156 genomes overlapping with gnomAD v2.1.1), the DiscovEHR dataset (11) (50,000 exomes) and UK Biobank (12) (200,000 exomes). Genomic coordinates of DiscovEHR and gnomAD v2.1.1 variants were converted from GRCh37 to GRCh38 reference assembly using LiftOver (13). Variants were then filtered to those single nucleotide variants (SNVs) passing each dataset's quality control filters, annotated using the Variant Effect Predictor (VEP) (Release 101) (14) for positions within Ensembl transcripts and consequence for filtering to synonymous and missense only.

Ensembl transcripts were downloaded from the Ensembl database (v101) (15) using the Bioconductor's biomaRt (16) package. RefSeq transcripts were downloaded from NCBI RefSeq (17) using the biomartr() (18) package for NM mRNA transcripts, NP coding sequences and

paired with Ensembl transcripts with identical Consensus CDS (CCDS) (19) sequence identifiers. A simulated set of all possible variants was generated by considering every possible single nucleotide substitution (3 variants per nucleotide in the sequence), filtered to missense and synonymous variants, and annotated using VEP to calculate the expected proportion of missense variants.

For validation purposes, ClinVar (20) missense variants were retrieved from the NCBI FTP server and subset based on their labels to pathogenic, likely pathogenic, benign and likely benign variants. Catalogue of Somatic Mutations in Cancer (COSMIC) v92 (21) variants were downloaded from their website and filtered to confirmed somatic missense variants. The FATHMM SwissProt/TrEMBLdisease variants dataset and FATHMM cancer-associated missense variants datasets were also retrieved for additional comparisons (22).

Sequence-based MTR scores can also be viewed in MTR3D, calculated using window sizes of 21, 31 and 41. MTR v1 refers to the MTR scores calculated using gnomAD v1 (23). MTR v2 refers to the current sequence-based MTR scores derived from variation from gnomAD v2 and v3, UK Biobank and DiscovEHR (7).

Calculation of the MTR scores across gene sequences

Missense Tolerance Ratio (MTR) scores were calculated using a sliding window of 21, 31 and 41 codons across each Ensembl and RefSeq transcript by comparing the observed proportion of missense variants to the expected proportion of variants (Equations 1-3).

For a given window $W_i^{H,J}$ and with selected window size w, the window is defined as:

$$\begin{aligned} &\text{where } i = \text{ residue position} \\ &H = \max\left(1, i - \tfrac{w-1}{2}\right) \\ &J = \min\left(\text{transcript length}, i + \tfrac{w-1}{2}\right) \end{aligned} \quad (1)$$

Within each window, the missense and synonymous variants are summed at each amino acid position y_i for both the observed and expected datasets (Equation 2), and the ratio is taken (Equation 3).

$$\begin{aligned} &y_i = \sum_{x_m \in \{W_i^{H,J}\}} x_m \\ &\forall x \in \{missense\_obs, synonymous\_obs, \\ &\quad missense\_exp, synonymous\_exp\} \end{aligned} \quad (2)$$

$$MTR_i = \frac{missense\_obs_i \,/\, (missense\_obs_i + synonymous\_obs_i)}{missense\_exp_i \,/\, (missense\_exp_i + synonymous\_exp_i)} \quad (3)$$

Alignment of transcripts to protein tertiary structures

UniProtKB's ID mapping table was used to identify pairings between Ensembl and RefSeq transcripts with their corresponding experimental and homology modelled PDB structures and chains (24). Experimentally determined protein structures were downloaded from RCSB Protein Data Bank (25), selecting only the primary biological assembly for each structure. Homology models of Ensembl or RefSeq transcripts were generated using SWISS-MODEL (26) and an in-house homology modelling pipeline using Modeller (27). We considered all potential templates with at least 30% identity for alignments longer than 100 residues and at least 70% identity for alignments shorter than 100 residues. Following minimization in Foldx, the quality of the homology models was assessed using Procheck (28), Molprobity (29) and WHATIF (30). The distribution of QMEAN Z-scores for the homology models was examined, revealing that over 76.9% of models have a Z-score above -4.0, indicating structural features of the homology models are comparable to what would be expected from high resolution X-ray structures (Figure S1).

Ensembl and RefSeq transcripts were aligned to protein tertiary structures in R. Transcripts, metadata and sequences were parsed using data.tables, PDB files were parsed using bio3d (31) and aligned using a ClustalW (32) pairwise alignment via the msa package (33). To take into consideration added and omitted residues (for example unresolved loops) and partial structures (where the experimental structure was generated from a region of the gene, for example a single domain), the alignment was then split where there were gaps of at least 3 residues. These were then considered separately for congruence of >50% and minimum length of 5 residues in order to allow unaligned regions to be discarded. 42,312 Ensembl transcripts and 32,845 RefSeq transcripts were aligned to 40,277 unique RCSB PDB structures, 41,861 unique SWISS-MODEL homology models and 43,477 unique homology models generated using Modeller.

Calculation of the MTR3D scores

Population variation and simulated data of all possible variants, as described above, were mapped to residues within the PDB structure files. At each residue position, in x,y,z coordinates in angströms, missense and synonymous variants were summed based on any residue within a distance of 5, 6 and 8 Å respectively. Proximal residues with at least one atom within each of these spheres were considered (Figure S2).

For a given window $W_i^{(x_1,x_2),(y_1,y_2),(z_1,z_2)}$ as a sphere of distance w, taken from the defined x, y, z coordinates of a residue (Equation 4),

where

$i =$ residue position

$$x_1 = x - w;\ x_2 = x + w$$
$$y_1 = y - w;\ y_2 = y + w \qquad (4)$$
$$z_1 = z - w;\ z_2 = z + w$$

Observed and expected missense and synonymous variants were summed for each window at each residue y_i (Equation 5).

$$y_i = \sum_{x_m \in \{W_i^{(x_1,x_2),(y_1,y_2),(z_1,z_2)}\}} x_m$$
$$\forall x \in \{missense\_obs, synonymous\_obs, missense\_exp, synonymous\_exp\} \quad (5)$$

$$MTR_i = \frac{missense\_obs_i \,/\, (missense\_obs_i + synonymous\_obs_i)}{missense\_exp_i \,/\, (missense\_exp_i + synonymous\_exp_i)} \quad (6)$$

The MTR3D was then calculated at each position, considering only positions with at least 3 observed variants (Equation 6).

MTR3D scores for both ClinVar and COSMIC missense variants were also compared at the different radii of 5 Å, 6 Å and 8 Å, and separately for experimentally determined and homology modelled structures (Figure S3 and S4). This revealed that the 5 Å window size was most informative.

Additionally, both the MTR and MTR3D were calculated for specific populations using a subset of the gnomAD database with sufficient representation of a given population (Admixed American (AMR), Non-Finnish European (NFE) and South Asian (SAS).

MTRX Consensus score

To assess the extent to which the combination of MTR and MTR3D scores are able to distinguish between pathogenic and non-pathogenic variants, a machine learning model was trained. Uniquely observed missense variants from ClinVar with no conflicting evidence of pathogenicity were assigned the class labels “P”, where clinical significance was denoted “Pathogenic” or “Likely pathogenic”, or “B” for “Benign” or “Likely benign”.

To develop the MTR consensus score, we considered the missense tolerance scores from MTR3D (using a radius of 5 Å), and the previous sequence-based scores from MTR v1 and MTR v2. The performance of each score was considered individually and in combination. In addition, general structural properties including measures of depth, residue solvent accessibility (RSA) and Psi/Phi angles at each residue position, calculated using DSSP 3.0 (34) and Biopython (35), were also considered.

Selecting the most informative features based on predictive performance (Table S1), a classifier was generated using Random Forest Classification (trees=100, depth=unlimited, number of features=unlimited) with the scikit-learn Python toolkit (36) and evaluated under 10-fold cross-validation, with the best performing model selected based on the Area Under the ROC curve (AUC) and Matthew’s Correlation Coefficient (MCC). The final classifier MTRX uses MTR3D,

MTR v2 21-codon windows, MTR v1 41-codon windows and RSA as evidence to distinguish between variant classes. Only positions with a valid score for these four metrics were given a consensus score.

WEB-SERVER

We have implemented MTR3D as a user-friendly and freely available web-server application (http://biosig.unimelb.edu.au/mtr3d). The server front end was developed using Materialize framework version 1.0.0, and the back end was built using Python 2.7 via the Flask framework (version 1.0.2). The web-server is hosted on a Linux Server running Apache2.

Input

MTR3D can be used to either browse a table of the full set of available gene transcript – PDB structure – chain pairings (Figure S5), or to search for a specific gene or transcript directly. Names are not case-sensitive.

On the viewer page (Figure 1) after making a selection, users may select alternate transcripts or alternate structures available for the current transcript or select between different distance calculations from a set of pre-computed options. Sequence-based MTR scores including population-specific MTRs can also be visualised directly on the structure. Users may also submit a protein position or list of protein residues to be highlighted on the structure, based on the transcript's protein position.

Output

A line graph using Bokeh is displayed to show the currently selected MTR scores as a 2D representation. This also provides a visualisation of which protein positions of the transcript are present in the currently viewed protein structure. Low scoring MTR3D scores indicate intolerance and purifying selection acting on that region, while high MTR3D scores indicate tolerance and, where MTR3D > 1.0, indicate that variation may be positively selected for in this region.

A viewer to interact with the protein structure is provided, displaying MTR scores mapped onto the structure, where blue coloured regions indicate tolerance and red regions indicate intolerance. The structure can be rotated, zoomed and translated. Individual residue information is displayed when hovering over the structure.

If residues have been selected, a red dot denotes their positions is highlighted on the line graph, and their side chains are displayed in stick format on the structure view.

Both the line graphs and structure representations can be downloaded as image files as currently displayed. A table of MTR scores with alignments between transcript protein positions and

structure residue numbers can also be downloaded as a CSV file, as well as the currently displayed PDB structure itself.

API

An Application Programming Interface (API) implementation is also available for the MTR3D scores for facilitating integration of our method with other systems and applications. Users may query an Ensembl transcript, RefSeq transcript, or HGNC symbol, and may optionally specify a protein position, specific PDB:chain identifier and specific score name. If no specific PDB:chain is supplied, the API will default to the structure with the most coverage for that transcript's alignment to the structure. If no protein position is supplied, the API will return all scores across the currently selected structure. If a specific score is selected, the API will only return values for that score. Results are returned as a JSON object.

Datasets

A bulk download is available via the web-server to download the full set of scores for Ensembl and RefSeq transcripts mapped to the experimental and homology structures. ClinVar disease variants, COSMIC somatic variants and DiscovEHR population control variants used for validation are also available for download via the web-server.

VALIDATION

Performance on disease-ascertained variants

MTR3D was assessed for its ability to differentiate pathogenic from non-pathogenic variants by comparing MTR3D scores across the ClinVar dataset. For each ClinVar gene transcript, a single protein structure with the greatest number of matching residues was selected, then ClinVar variants were filtered to uniquely observed variants by removing duplicate observations in order to prevent bias towards gene symbols with many transcripts or overrepresented variants. Note that validation could only be performed on ClinVar genes with a valid structure (2,752 experimental structures, 6,333 homology modelled structures). Performance of experimentally determined protein structures was assessed separately to the homology modelled structures to assess whether both show similar enrichment of pathogenic variants within intolerant regions (Figure S3).

Intolerant regions were found to be significantly enriched for ClinVar non *de novo* pathogenic variants (n = 14,547) and *de novo* pathogenic variants (n = 725) than benign variants (n = 7,086) for both experimentally determined and homology modelled structures (Figure 2A; Mann-Whitney U test $p < 2.2 \times 10^{-16}$ and $p < 2.2 \times 10^{-16}$, respectively). At a MTR3D (5 Å) less than 0.5, which we consider to be intolerant, 537 of 725 ClinVar *de novo* pathogenic and 5,030 of 14,547 ClinVar non *de novo* pathogenic variants were observed, while only 856 of 7,086 benign variants were found

here. The MTR3D scores was further assessed using the FATHMM SwissProt/TrEMBL training dataset and found to perform similarly (Mann-Whitney U test $p < 2.2 \times 10^{-16}$).

Performance on cancer-ascertained variants

COSMIC unique somatic missense variants from cancer samples were compared with DiscovEHR population variants to determine whether there is significant enrichment of COSMIC variants within intolerant regions compared with standing variation within the population (Figure 2B). We defined a proposed cutoff of 0.75 to denote intolerance, however the ideal cutoff will vary depending on the gene under investigation. Over two thirds of COSMIC variants (18,981 / 27,520) were found to have a MTR3D below 0.75. A significant enrichment was found in both experimentally determined and homology models for COSMIC variants (Figure S4; Mann-Whitney U test $p < 2.2 \times 10^{-16}$ and $p < 2.2 \times 10^{-16}$, respectively). Using the FATHMM cancer-associated training dataset, we find similar enrichment for cancer-associated variants within intolerant regions (Mann Whitney U-test $p < 2.2 \times 10^{-16}$).

Interestingly, when we compared the intolerance scores of variants in tumour suppressor (n = 116 genes) and oncogenes (n = 91 genes) separately, while background control variation did not reveal any significant difference, cancer-ascertained variants in oncogenes were associated with significantly lower MTR3D scores than those in tumour suppressors (Figure S6). This is likely due to the larger effect of purifying selection of dominant variants.

Performance of the MTRX consensus score

A consensus score, MTRX, was built using the MTR3D scores, together with sequence-based MTR scores and general structural properties, using the ClinVar database (n = 23,406 variants). The MTRX represents a likelihood of a variant being pathogenic [0-1]. The distribution of MTRv1, MTRv2, MTR3D and RSA across the dataset shows clear differences between benign and pathogenic variants (p-value < 0.0001, Figure S7), and interestingly there is not a strong correlation between the spatial and sequence based scores (Figure S8). The overlap in intolerant regions between the spatial and sequence based scores, indicated that while there was significant agreement, over 18% of the intolerant regions under selective pressure were identified by only the spatial based scores, in particular in sequence based windows with limited data (Figure S9).

Table 1 shows the predictive performance of individual scores and their combination under 10-fold cross validation. Individually, MTR scores achieved AUCs of 0.63 (MTR3D; 5 Å), 0.64 (MTR v2; 21 codons) and 0.67 (MTR v1; 41 codons), respectively (Figure 2D). While individual features only presented a modest ability of distinguishing between pathogenic and benign variants, a significant improvement in performance (p-value < 0.001) is observed when scores are combined in a consensus, achieving an AUC of 0.85 and MCC of 0.49, demonstrating the complementary nature of the scores. Performance is further improved when structural properties (residue relative solvent accessibility) is also considered (Figure 2D; AUC of 0.90 and MCC of 0.62). An analysis of feature importance also showed a high level of complementarity between MTR scores and the selected structural property (Figure S10).

Exploring the learned trees further, we observe that the top of the majority of the decision trees uses as first branching point an RSA of 20.7% (Figure 2C). Interestingly, this is consistent with general thresholds for considering residues as either buried (RSA <20%) or exposed (RSA >20%) (37,38). For buried residues, MTRX considered a variant pathogenic if the MTR3D score was below 0.73 or the MTRv1 score below 0.68 (Figure 1A). For exposed residues, variants were considered pathogenic if their MTR3D score was below 0.58, indicating the need for stronger evidence of intolerance to label exposed residues as pathogenic than buried residues. These two simple rules covered over a quarter of the data.

CONCLUSION

The MTR3D application provides an intuitive and programmable interface to explore intolerance and purifying selection within protein tertiary structures and across the flat sequence. The MTR3D and MTR consensus scores are versatile metrics to assess the likelihood of pathogenicity in patient-ascertained variants, as well as measures to identify novel important regions within protein structures that may be overlooked by traditional metrics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

C.H.M.R is funded by a Melbourne Research Scholarship. D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MR/M026302/1); and the Jack Brockhoff Foundation (JBF 4186, 2016). D.B.A., M.S. and D.E.V.P. were funded by an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). Supported in part by the Victorian Government's Operational Infrastructure Support Program.

REFERENCES

1. Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. and Ng, P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, **40**, W452-457. http://dx.doi.org/10.1093/nar/gks539
2. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248-249. http://dx.doi.org/10.1038/nmeth0410-248

3. Ghosh, R., Oak, N. and Plon, S.E. (2017) Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol*, **18**, 225. http://dx.doi.org/10.1186/s13059-017-1353-5
4. Epi25 Collaborative. Electronic address, s.b.u.e.a. and Epi, C. (2019) Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. *Am J Hum Genet*, **105**, 267-282. http://dx.doi.org/10.1016/j.ajhg.2019.05.020
5. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. and Goldstein, D.B. (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*, **9**, e1003709. http://dx.doi.org/10.1371/journal.pgen.1003709
6. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M. and Daly, M.J. (2017) Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353. http://dx.doi.org/10.1101/148353
7. Silk, M., Petrovski, S. and Ascher, D.B. (2019) MTR-Viewer: identifying regions within genes under purifying selection. *Nucleic Acids Res*, **47**, W121-W126. http://dx.doi.org/10.1093/nar/gkz457
8. Iqbal, S., Perez-Palma, E., Jespersen, J.B., May, P., Hoksza, D., Heyne, H.O., Ahmed, S.S., Rifat, Z.T., Rahman, M.S., Lage, K. *et al.* (2020) Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc Natl Acad Sci U S A*, **117**, 28201-28211. http://dx.doi.org/10.1073/pnas.2002660117
9. Wagih, O., Galardini, M., Busby, B.P., Memon, D., Typas, A. and Beltrao, P. (2018) A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol Syst Biol*, **14**, e8430. http://dx.doi.org/10.15252/msb.20188430
10. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434-443. http://dx.doi.org/10.1038/s41586-020-2308-7
11. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C.V., Staples, J., Gonzaga-Jauregui, C. *et al.* (2016) Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, **354**. http://dx.doi.org/10.1126/science.aaf6814
12. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, **12**, e1001779. http://dx.doi.org/10.1371/journal.pmed.1001779
13. Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N. *et al.* (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*, **47**, D853-D858. http://dx.doi.org/10.1093/nar/gky1095
14. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol*, **17**, 122. http://dx.doi.org/10.1186/s13059-016-0974-4
15. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res*, **48**, D682-D688. http://dx.doi.org/10.1093/nar/gkz966
16. Durinck, S., Spellman, P.T., Birney, E. and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, **4**, 1184-1191. http://dx.doi.org/10.1038/nprot.2009.97
17. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2016) Reference sequence (RefSeq)

database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, **44**, D733-745. http://dx.doi.org/10.1093/nar/gkv1189
18. Drost, H.G. and Paszkowski, J. (2017) Biomartr: genomic data retrieval with R. *Bioinformatics*, **33**, 1216-1217. http://dx.doi.org/10.1093/bioinformatics/btw821
19. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, **19**, 1316-1323. http://dx.doi.org/10.1101/gr.080531.108
20. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, **46**, D1062-D1067. http://dx.doi.org/10.1093/nar/gkx1153
21. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E. *et al.* (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, **47**, D941-D947. http://dx.doi.org/10.1093/nar/gky1015
22. Shihab, H.A., Gough, J., Mort, M., Cooper, D.N., Day, I.N. and Gaunt, T.R. (2014) Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics*, **8**, 11. http://dx.doi.org/10.1186/1479-7364-8-11
23. Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J. and Petrovski, S. (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res*, **27**, 1715-1729. http://dx.doi.org/10.1101/gr.226589.117
24. UniProt, C. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, **49**, D480-D489. http://dx.doi.org/10.1093/nar/gkaa1100
25. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242. http://dx.doi.org/10.1093/nar/28.1.235
26. Bienert, S., Waterhouse, A., de Beer, T.A., Tauriello, G., Studer, G., Bordoli, L. and Schwede, T. (2017) The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res*, **45**, D313-D319. http://dx.doi.org/10.1093/nar/gkw1132
27. Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779-815. http://dx.doi.org/10.1006/jmbi.1993.1626
28. Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283-291. http://dx.doi.org/doi:10.1107/S0021889892009944
29. Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, **66**, 12-21. http://dx.doi.org/10.1107/S0907444909042073
30. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J Mol Graph*, **8**, 52-56, 29. http://dx.doi.org/10.1016/0263-7855(90)80070-v
31. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695-2696. http://dx.doi.org/10.1093/bioinformatics/btl461
32. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680. http://dx.doi.org/10.1093/nar/22.22.4673

33. Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C. and Hochreiter, S. (2015) msa: an R package for multiple sequence alignment. *Bioinformatics*, **31**, 3997-3999. http://dx.doi.org/10.1093/bioinformatics/btv494
34. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637. http://dx.doi.org/10.1002/bip.360221211
35. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422-1423. http://dx.doi.org/10.1093/bioinformatics/btp163
36. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, **12**, 2825-2830.
37. Savojardo, C., Manfredi, M., Martelli, P.L. and Casadio, R. (2021) Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences. *Frontiers in Molecular Biosciences*, **7**. http://dx.doi.org/10.3389/fmolb.2020.626363
38. Chen, H. and Zhou, H.-X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Research*, **33**, 3193-3199. http://dx.doi.org/10.1093/nar/gki633

FIGURES

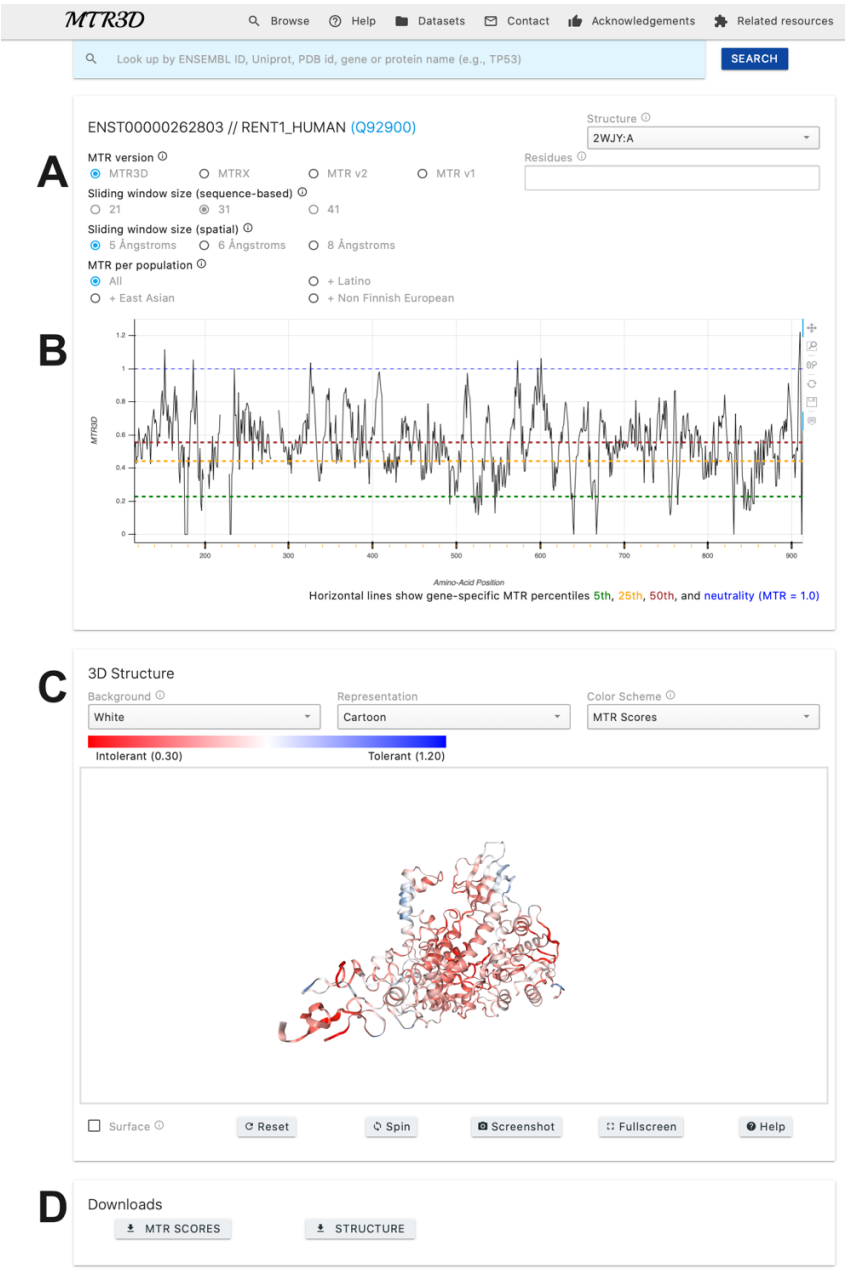


Figure 1. MTR3D viewer page. (A) Users may select between different structures and sequence-based, spatial-based and consensus scores for the currently selected transcript. Users may also select between window sizes and population estimates. (B) Line graph showing the alignment of scores to the currently selected transcript and structure. Gaps in the plot indicate regions not congruent or not present in the protein tertiary structure. Horizontal lines indicate MTR percentiles for the current transcript at 5th, 25th, 50th and MTR = 1. (C) The selected protein structure is displayed and coloured by the currently selected MTR score, where red and blue represent

intolerance and tolerance respectively. (D) Download links for the MTR scores for the currently selected structure or the currently shown PDB.

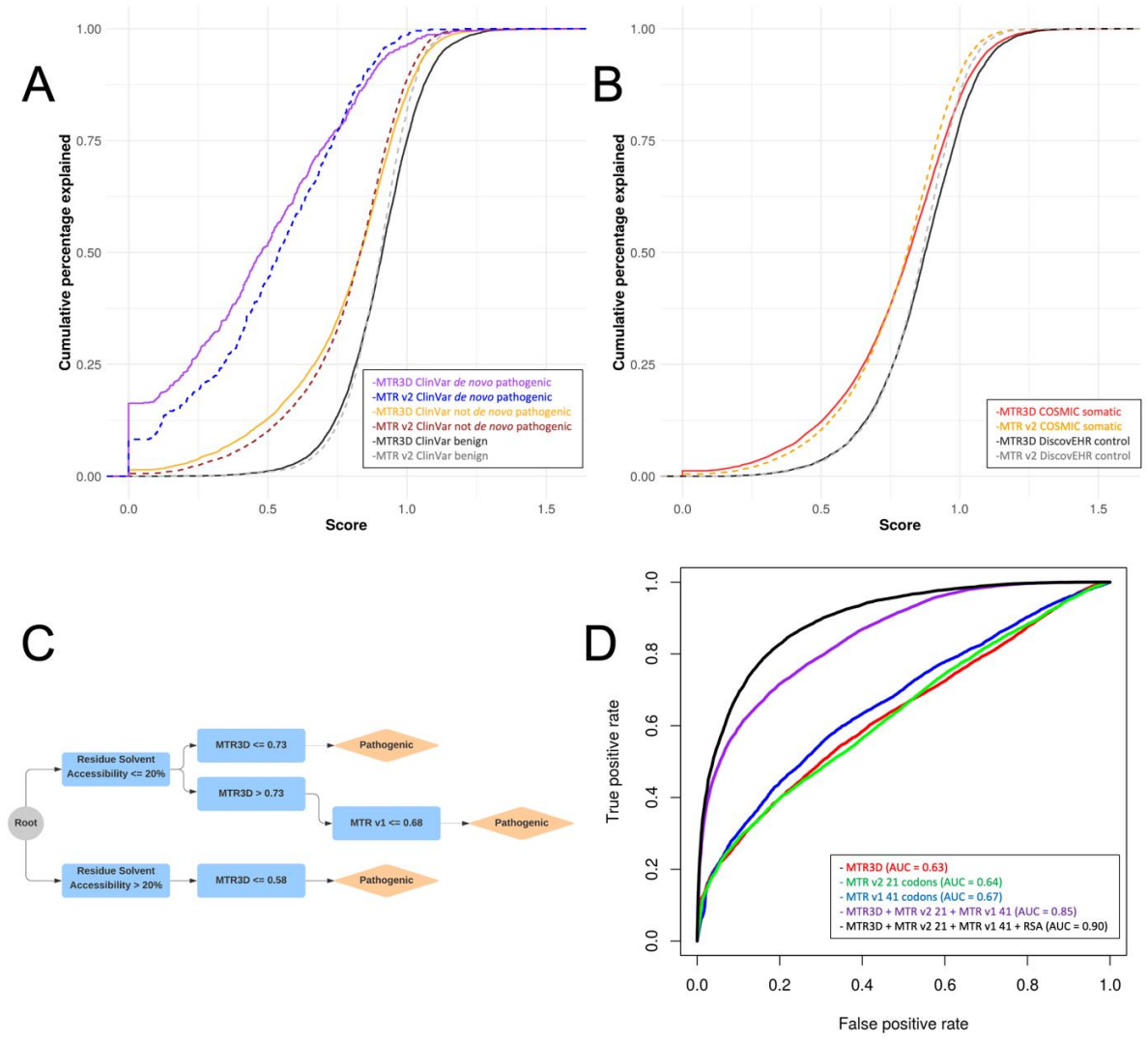


Figure 2: Performance of MTR3D and consensus score on identification of disease and cancer-ascertained variants. Comparison of the spatial- and sequence- based MTR scores using disease-associated variants. (A) Cumulative distribution graph comparing MTR3D (5 Å) and MTR v2 (31 codons) in ClinVar *de novo* pathogenic missense variants (purple, blue respectively), ClinVar not *de novo* pathogenic missense variants (orange, brown respectively) and ClinVar benign missense variants (black, grey respectively). (B) Cumulative distribution graph comparing COSMIC somatic missense variants MTR3D (5 Å) scores (red), MTR v2 (31 codons) scores (orange), with DiscovEHR population missense variants observed within the same genes (black, grey respectively). (C) Decision tree representation of the most informative scores used in the generation of the consensus metric calculated using a Random Forest model. Cut-offs were determined based on 10-fold cross-validation. (D) Area under the Curve (AUC) plot showing classification specificity/sensitivity for MTR3D (5 Å) (red), MTR v2 21 codons (green), MTR v1 41 codons (blue), MTR consensus using MTR3D (5 Å) + MTR v2 21 + MTR v1 41 (purple), and with RSA included (black).

TABLES

Table 1: Predictive performance of MTRX consensus scores on ClinVar variants

Score	TP rate	FP rate	Precision	Recall	AUC	MCC
MTR3D 5 Å	0.64	0.57	0.60	0.64	0.63	0.10
MTRv2 (21 codons)	0.64	0.55	0.60	0.64	0.64	0.12
MTRv1 (41 codons)	0.65	0.49	0.63	0.65	0.67	0.17
MTR3D + MTRv2 + MTRv1	0.77	0.30	0.77	0.77	0.85	0.49
MTRX	0.83	0.22	0.83	0.83	0.90	0.61

Appendix E

Machine Learning of ECG Waveforms to Improve Selection for Testing for Asymptomatic Left Ventricular Dysfunction

Guest Editor: Alan Fraser, MD

Machine Learning of ECG Waveforms to Improve Selection for Testing for Asymptomatic Left Ventricular Dysfunction

Short title: ECG machine-learning and LV dysfunction

Elizabeth L. Potter MBBS BSc[1,2], Carlos H. M. Rodrigues BSc[1,3], David B. Ascher PhD[1,3], Walter P. Abhayaratna MBBS PhD[4,5], Partho P. Sengupta MD[6], Thomas H. Marwick MBBS, PhD, MPH[1,2]

Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia[1]; School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia[2]; Melbourne University, Melbourne Australia[3], Australian National University Medical School, The Australian National University, Canberra, Australian Capital Territory, Australia[4]; Division of Medicine, Canberra Hospital, Canberra, Australian Capital Territory, Australia[5]; West Virginia University Heart and Vascular Institute, Morgantown, West Virginia, USA[6]

Corresponding author:
Prof Tom Marwick
Baker Heart and Diabetes Institute
75 Commercial Road
Melbourne, Vic 3004, Australia
tom.marwick@baker.edu.au
5120 text words including references, 40 references, 3 figures, 4 tables, supplemental material

Funding: The work was partially supported by a Partnership grant (1149692) from the National Health and Medical Research Council, Canberra, the Ian Potter Foundation, Melbourne, and the Baker Heart and Diabetes Institute. Dr Potter is supported by a Monash University postgraduate scholarship.

Conflict of Interest: No authors report conflicts of interest.

Authors contributions: ELP and THM contributed to the conception and design of the work. ELP, THM, WPA and PPS contributed to the acquisition of the data. CR and DA contributed to the analysis of the data. ELP drafted the manuscript. All authors critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of work ensuring integrity and accuracy.

Abstract

Background. Asymptomatic left ventricular dysfunction (LVD) has management implications, but routine echocardiography is not undertaken in subjects at risk of heart failure. Signal processing of the surface ECG using continuous wave transforms (CWT) can identify abnormal myocardial relaxation. We sought whether machine-learning from CWT-processed "energy waveform" ECG (ewECG) could be integrated with echocardiographic assessment of subclinical systolic and diastolic LVD.
Methods. EwECG and echocardiography were undertaken in 398 participants at risk of HF. Reduced global longitudinal strain (GLS≤16%)), diastolic abnormalities (E/e'>15, left atrial enlargement with E/e'>10 or impaired relaxation) or LV hypertrophy defined LVD. EwECG feature selection and supervised machine-learning by Random Forest (RF) classifier was undertaken with 643 CWT-derived features and the Atherosclerosis Risk in Communities (ARIC) heart failure risk score.
Results. The ARIC score and 18 CWT features were selected to build a RF predictive model for LVD in a training data set (n=287, 60% female, median age 71(68-74) years). Model performance was tested in an independent group (n=111, 49% female, median age 61(59-66) years), demonstrating 85% sensitivity and 72% specificity (AUC 0.83,95% CI 0.74-0.92). With ARIC score removed, sensitivity was 88% and specificity, 70% (AUC 0.78,95% CI 0.70-0.86). RF models for reduced GLS and diastolic abnormalities including similar features had sensitivities that were unsuitable for screening. Conventional candidates for LVD screening (ARIC score, NT-proBNP and standard automated ECG analysis) had inferior discriminative ability. Integration of EwECG in screening of people at risk of HF would reduce need for echocardiography by 56% while missing 12% of LVD cases.
Conclusion. Machine-learning applied to ewECG is a sensitive screening test for LVD and its integration into screening of patients at risk of HF would reduce the number of echocardiograms by over half.

Keywords: Left ventricular dysfunction, electrocardiography, machine-learning, screening

Abbreviations:
ARIC - Atherosclerosis Risk in Communities
CAD - coronary artery disease
cNRI - continuous net reclassification improvement
CWT - continuous wave transforms
EwECG - energy waveform ECG
GLS - global longitudinal strain
HF - heart failure
IDI - integrated discrimination improvement
LVD – left ventricular dysfunction
RF - Random Forest

Introduction

The echocardiographic recognition of structural and functional cardiac abnormalities among patients with HF risk factors identifies asymptomatic LV dysfunction (LVD), and thereby guides therapy (1). Despite this, routine echocardiography is rarely performed in people with HF risk factors, other than in coronary artery disease (CAD) (2). No viable screening test has evolved to better direct selection for echocardiography.

Standard 12-lead electrocardiography (ECG) is a potential initial investigation in people with HF risk factors. A range of ECG abnormalities, e.g. arrhythmias, conduction disturbances and voltage patterns, may relate to underlying LV dysfunction. The potential for a form of signal processing, known as continuous wavelet transform (CWT), to reveal abnormalities in a standard ECG signal has been recognized for over 20 years (3,4). However, only recently have patterns in CWT-processed ECG signals or 'energy waveform ECG (ewECG)' been demonstrated to predict functional LV abnormalities, specifically abnormal relaxation (5,6). Recording an ewECG requires no additional time or expertise and simultaneously displays a standard 12-lead ECG trace, making it a feasible test for use in the community. However, although there is an association between repolarization measures and abnormal myocardial relaxation (7), whether prior findings in populations presenting for echocardiography (6) extend to the detection of asymptomatic *systolic* LV dysfunction in community populations at risk of HF is unknown. Indeed, the most appropriate use of this technology would be to guide definitive echocardiographic assessment as part of a screening process. We hypothesized that machine-learning algorithms applied to ewECG data could identify LVD in a community population at risk of HF, and that depolarization and repolarization CWT features were associated with systolic and diastolic dysfunction (DD) respectively.

Methods

Study participants. Participants were recruited from the community as part of the ongoing Victorian Study of Echocardiographic detection of Left ventricular dysFunction (Vic-ELF; ACTRN 12617000116325), in Melbourne, Australia. Participants were aged ≥65 years with at least one of the following HF risk factors; obesity (BMI ≥30 kg/m^2), type 2 diabetes mellitus or hypertension (systolic blood pressure ≥140mmHg or on medication). Exclusion criteria included LV ejection fraction (LVEF) ≤40%, known symptomatic heart failure (or diagnosed at baseline screening), known coronary artery disease (CAD; excluded due to the routine use of echocardiographic assessment in this group), moderate or greater valvular heart disease, renal impairment and symptoms of HF.

The testing dataset comprised an analogous group of 111 prospectively recruited asymptomatic people without established cardiovascular disease, with the same inclusion and exclusion criteria, in Canberra (Australian National University Medical School), Australia (n=79); New York, NY (Icahn School of Medicine), and Morgantown, WV (University of West Virginia), USA (n=32). This geographical heterogeneity aimed to test model generalizability. The relevant institutional review boards approved the study and participants gave written informed consent.

Clinical measures. Baseline measures and procedures pertinent to this sub-study included: body mass index, resting averaged systolic blood pressure (SBP) and diastolic blood pressure (DBP), heart rate (HR), documentation of cardiovascular risk factors, comorbidities and medications. Clinical data were used to calculate the 4 year risk of incident symptomatic HF using the Atherosclerosis Risk in Communities (ARIC) HF risk score, which has demonstrated utility in risk stratification in subclinical HF (8). Biochemical markers of renal function and N-terminal pro-brain natriuretic peptide (NT-proBNP) were also measured.

Standard electrocardiography and energy waveform ECG. After standard ECG lead placement, subjects underwent ewECG evaluation using a device that is CE-marked but not approved for use in the USA (MyoVista Version 2.0, HeartSciences, Southlake, TX). The MyoVista ewECG interface displays a standard 12-lead, ECG trace as well as an automated diagnostic interpretation based on the University of Glasgow 12-lead ECG interpretive analysis algorithm, which provides both quantitative parameters and qualitative interpretations (9,10). The ECG signal is deconstructed and presented graphically in an energy scalogram (red [high energy] to blue [low energy]) depicting an energy distribution by time (x-axis) and frequency (y-axis) (the "energy waveform") (Figure 1). As described previously (5), energy is expressed as coefficients reflecting agreement between wavelet and signal at varying scales, rather than a discrete energy measurement (11). A total of 643 CWT features (energies, frequencies and ratios) at defined points in the cardiac cycle are generated by proprietary software throughout the cardiac cycle. As our prior work demonstrated the existing automated interpretative algorithms to be insufficiently sensitive to detect LVD (12), we used the complete CWT output.

Echocardiography. On the same day as ewECG a transthoracic 2-D and Doppler echocardiographic study was performed using standard equipment (ACUSON SC2000, Siemens Healthcare USA, Mountain View, CA) and transducer (4V1c, 1.25 to 4.5 MHz; 4Z1c, 1.5 to 3.5 MHz) in accordance with American Society of Echocardiography guidelines. LV systolic function was assessed by global longitudinal strain (GLS) computed using speckle-tracking (Syngo VVI, Siemens Healthcare USA, Mountain View, CA). GLS was the average of regional strains in the apical 2-chamber, 4-chamber and long axis views. Diastolic function was assessed by measuring mitral inflow peak early diastolic velocity (E), peak late diastolic velocity (A), E/A ratio, septal and lateral mitral annular early diastolic velocities (e') and the E/e' ratio. LA volume index (LAVi) was calculated from maximal LA volume using

biplane images and indexed to body surface area; LA enlargement was defined as LAVi >34 ml/m^2. Left ventricular hypertrophy (LVH) was defined as LVMi >95 g/m^2 in women and >115 g/ m^2 in men. LVD was defined by either i) abnormal structure (LVH), ii) abnormal GLS $\leq$16%, or borderline GLS (17-18%) with impaired relaxation (IR) or left atrial enlargement (LAE), iii) diastolic dysfunction (E/e' >15 or E/e'>10 with LAE or IR with LAE).

We developed predictive models for: a) LVD, b) systolic dysfunction (GLS$\leq$16%), c) diastolic dysfunction.

Machine-learning classification model. A supervised machine-learning approach was used to predict LVD status. We used the Random Forest (RF) classifier algorithm in the module Scikit-learn (Python Software Foundation, https://www.python.org/) (13). All hyperparameters for the algorithm were entered according to the library's documentation and have been summarised in Table 5 in the Supplemental material. Given the high dimensionality of the data, including the fact that ewECG features were more numerous than subjects, we undertook a process of feature selection to identify those features with the most predictive information relevant to LVD (Supplemental material, Figure 1). All CWT features plus the ARIC HF risk score were offered in feature selection. The ARIC score was included, so as to evaluate importance of this easily attainable clinical variable against ewECG.

The feature selection approach evaluated the performance of all individual features (*N*), using area under the ROC curve, and selected the best performing in the first round. Then, each of the remaining (*N-1*) features was paired with the selected one to identify the pair that gave the best performance. This process was repeated until all features were selected and the combination of features that provided the best performance was ascertained. Other approaches were tested but were less successful (Supplemental material Table 1). We also evaluated feature importance, extracted from the output of Random Forest (which uses Gini

importance). This indicates the contribution made by each feature to the model's predictive performance. Practically, higher importance means that the decision-making error associated with a feature in the nodes across all decision trees in the forest is less than using other features in other nodes.

A 5-fold cross validation on the training data set was used to internally validate model performance with subsequent external validation on the separate/test dataset. The output of the RF model is a continuous probability score with a threshold of 50% for dichotomising predicted outcome e.g. LVD vs. no LVD. When evaluating the performance on the external dataset, modification of this probability threshold was investigated to see if performances could be optimised. Unless otherwise stated, the 50% threshold was found to be optimal.

Statistical analysis. Continuous data are presented as mean ± standard deviation (SD) or median and interquartile range (IQR) depending on distribution after visual assessment. Between group differences for categorical data were tested using Pearson's chi-square, and for continuous variables the independent t-test or Wilcoxon rank-sum test was used depending on normality of distribution. Because the most important characteristic of a screening test is high sensitivity, cut-points were selected with the minimal number of false positives at a sensitivity closest to 90%. Discriminatory performance predictive models were assessed using area under receiver operating characteristic curve (ROC AUC). To evaluate the incremental utility of the machine learning models compared with the ARIC HF risk score as a base model, continuous net reclassification improvement (cNRI) and integrated discrimination improvement (IDI) were calculated. CNRI measures improvements in probabilities within events (i.e. increased probability) and non-events (i.e. decreased probabilities), with the addition of, in this case, the machine-learning models (14). IDI reflects the difference in discrimination slopes (probabilities for events minus non-events) between 2 models and is reported herein as the absolute IDI (15). For all analyses, statistical

significance was defined as a two-tailed p value <0.05. Analyses were conducted using STATA 15.1 (StataCorp, College Station, TX).

Results

Participants: Overall we included 398 participants (57% female, median age 69 (66-73) years) and of these, 171 (43%) had LVD. Baseline characteristics by HF stage are shown in Table 1. Compared with people with only risk factors, LVD was associated with older age, a higher proportion of hypertension, T2DM, increased heart rate and systolic blood pressure, and higher ARIC HF risk score. The proportion of an 'abnormal' Glasgow ECG analysis summary was 15% and 36% for people with risk factors and LVD, respectively (p<0.001). All echocardiographic measures differed significantly between HF stages.

Prediction of LVD by conventional methods. The ARIC HF risk score had an AUC of 0.72 (95% CI 0.67-0.77) for LVD discrimination. An optimized cut-point for sensitivity was identified as an ARIC HF risk score of 2.6, providing 90% sensitivity and 40% specificity. Similarly, the AUC for NT-proBNP was 0.53 with an optimized cut-off of 21pg/ml providing a sensitivity of 88% and specificity of 14%. Lastly, an abnormal ECG by Glasgow analysis had a sensitivity and specificity of 36% and 85%, respectively. In those with available NT-proBNP, adding NT-proBNP to ARIC HF risk score (AUC 0.65 [95% CI 0.59-0.71]) did not significantly improve discriminatory ability vs. ARIC alone (AUC 0.63 [95% CI 0.56-0.69], p=0.18). Furthermore, the addition of both NT-proBNP and abnormal ECG by Glasgow analysis did not significantly improve discriminatory ability (AUC 0.67 [95% CI 0.61-0.74], p=0.06) (Supplemental material, Figure 2).

Prediction of LVD by Random Forest classifier using ewECG. Of the 398 subjects, 287 (72%) were used to train the RF prediction model and 111 (28%) were used to test model performance. Compared with the training dataset, subjects in the test dataset were

significantly younger and there was a lower proportion of females, as well as participants with hypertension and diabetes. Furthermore, SBP, DBP and ARIC HF risk were significantly lower. The prevalence of the LVD composite was 23% in the test dataset compared to 51% in the training dataset ($p<0.001$) and a similar pattern was observed for abnormal GLS and diastolic abnormalities (Table 2).

The ARIC HF risk score was selected during feature selection along with 18 CWT features to train an RF model (Table 3). At a probability threshold of 0.51 (optimized for sensitivity), the sensitivity and specificity of the model for prediction of LVD on the test dataset were 85% and 72%, respectively (ROC AUC 0.83 (95% CI 0.74-0.92) (Table 4). With ARIC removed from the model, the optimized sensitivity and specificity for detection of LVD were 88% and 70%, respectively (ROC AUC 0.78 (95% CI 0.67-0.88), $p=0.32$ for difference between models) (Table 3, Figure 2). For a prevalence of 43%, this corresponded to a negative predictive value of 89% (95% CI 78-95%) and positive predictive value of 69% (95% CI 60-77%).

Incremental improvements in prediction were seen for both RF models compared with the ARIC HF risk score alone, as assessed by cNRI and IDI. For the RF model incorporating ARIC, cNRI was 0.79 (95% CI 0.23-1.17) and IDI 0.09 (95% CI 0.012-0.24). For the model incorporating only ewECG features cNRI was 0.94 (95% CI 0.46-1.29) and IDI 0.11 (0.017-0.255).

The RF classifiers were inspected to reveal their node features. For the LVD predictive model, features were temporally associated with both depolarization and repolarization and included several features derived from the energy/power spectrum i.e. certain ratios of harmonics within the power spectrum throughout cardiac cycles (Table 3).

Prediction of low global longitudinal strain by ewECG, using the Random Forest classifier. When the features from the predictive model for LVD were used to train a

predictive model for low GLS (≤16%), this was not able to identify any cases of low GLS (supplemental material, Table 2). After repeating feature selection, 16 features were found to confer peak predictive power (supplemental material, Table 3), and performance on the test dataset showed a sensitivity of 57% and a specificity of 90% (Table 4). However, the proportion with low GLS in the test dataset (6%) was significantly lower than the training dataset (19%, p=0.002), which is of significance in interpreting model performance. The CWT features selected for the low GLS model were predominantly power spectrum and repolarization features (ARIC HF risk score was not selected). Nine out of the 16 features were chosen for the LVD model.

Prediction of diastolic abnormalities by ewECG, using the Random Forest classifier. Overall, 91 (23%) exhibited diastolic abnormalities, with a significantly lower proportion in the test versus training dataset (13% vs. 27%, p=0.002, respectively [Table 2]). Again, features from the LVD predictive model trained for diastolic abnormalities were unable to discriminate (Supplemental material, Table 2). After repeated feature selection, a model with 14 features produced a sensitivity of 50% and a specificity of 90% (Table 4). Features selected for inclusion in the model included power spectrum and repolarization features as well as one depolarization-related features that also occurred in the LVD model (ARIC HF risk score was not selected) (supplemental material, Table 4). Ten out of the 14 features were also chosen for the LVD model.

Impact of ewECG on a screening process for LVD. On the basis of greatest sensitivity for LVD identification, the RF ewECG model without the ARIC HF risk score was considered most optimal. Use of this ewECG-based model to select people for echocardiography could reduce the number of echocardiograms by 56% ([true negatives + false negatives]/total screened x 100). However, 12% of cases of LVD would be missed. Alternatively, the RF model incorporating ARIC score and ewECG, would result in a 60% reduction in

echocardiograms but would miss 16% of cases of LVD. By comparison, using the best of the conventional methods i.e. ARIC HF risk score ≥2.6, echocardiograms would be reduced by only 27% and 11% of LVD cases would be missed (Central illustration).

Discussion

The application of signal processing (CWT), to a conventionally acquired ECG signal provides a viable screening test that can be integrated with echocardiographic screening for LV dysfunction. In our cohort of people at-risk of HF, our machine-learning (RF) algorithm provided 88% sensitivity and 70% specificity for detection of LVD, outperforming a clinical risk score, biomarkers and an established automated ECG analysis algorithm. Furthermore, we highlighted that CWT features required for identification of LVD differed slightly according to LV abnormality e.g. reduced systolic function or diastolic abnormality. The best-performing model was for prediction of a composite measure of LVD. There was no significant difference between a machine-learning model that incorporated clinical information i.e. the ARIC HF risk score, compared with ewECG features alone. If implemented in combination with echocardiography as a screening test, ewECG could reduce echocardiograms by 56% in screening for LVD. This is important given the 82% prevalence of subclinical HF in the community in those over 67 years (16). For the United States this would mean approximately 40 million people would be eligible for screening and ewECG could reduce the number of echocardiograms by around 23 million (17). Even if such a screening program did not eventuate, ewECG may serve as a gatekeeper to the echocardiography lab in high risk but asymptomatic individuals.

Electrocardiography and myocardial dysfunction. Structural and metabolic cardiac pathology manifesting as electrocardiographic abnormalities is well accepted. However, abnormalities may be too subtle for either the human reader or standard analytics to detect

(9). Accordingly, a recent study used artificial intelligence (AI) (convolutional neural networks) applied to standard ECG digital data to predict LVD (defined as LV ejection fraction ≤ 35%) with a sensitivity of 89%, specificity of 83% and AUC of 0.93 (18). Interestingly, those with a "false positive" result were 4 times as likely to develop LVD during 4-year follow-up, suggesting the AI could recognize early abnormalities. Another approach extracted advanced ECG (A-ECG) parameters (3D ECG parameters, QRS/T wave complexity parameters) including variability analysis (5-minute high fidelity ECG recording) as well as conventional ECG measures to devise a prediction score for myocardial disease (19). The authors demonstrated that a 5-parameter A-ECG score (derived using a feature selection technique and logistic regression) had 83% sensitivity and 93% specificity for LVD (LVEF<50%) in a group of predominately male subjects with either CAD or LVH. Interestingly, none of the features used in this score required an extended duration, high sample rate recording, and as such could be attained from a conventionally recorded ECG, after advanced analytics. These works and ours, demonstrate a growing body of evidence supporting the feasibility of electrocardiographic identification of LVD.

Continuous wavelet transforms processing and cardiac disease. Wavelet transforms have been applied to the ECG signal for measurement of intervals, noise reduction and importantly identification of abnormalities (20). One of the first applications was identification of ventricular late potentials (VLPs), microvoltage deflections after (and sometimes within) the QRS complex that are often obscured by noise (4,20). The detection of VLPs using CWT improved prediction of post-infarction ventricular arrhythmias from 52-72% and 64-76% for inferior and anterior infarctions, respectively, compared to standard signal filtering (21). Wavelet transforms have also revealed electrical similarities (abnormal frequency content) within the QRS complex in congenital and acquired long QT syndrome, providing insight into shared electro-pathophysiology (22).

While CWT-processed ECG has demonstrated value by revealing known or suspected electrical abnormalities, there are limited data on a direct association between CWT-processed ECG features and cardiac *function*. Associations between standard ECG features and cardiac dysfunction has focused on long QT syndrome, which is associated with increased isovolumetric relaxation time, altered tissue Doppler velocity profiles and mechanical dispersion (23,24). Furthermore, the interval from T-wave peak to T-wave end (TpTe) is increased in DD assessed by mitral inflow and tissue Doppler velocities (7). At the molecular level, DD is partly related to low amplitude calcium transients secondary to reduced calcium uptake into, and leakage from, the sarcoplasmic reticulum (25,26). Given calcium is a key modulator of the action potential duration, disturbances in the electrical signal on the surface ECG may be apparent in LV dysfunction. It follows that detailed decomposition of the ECG signal from a diseased myocardium may reveal characteristic abnormalities, and indeed, the machine-learning model using only ewECG features provided accurate detection of LVD. Recently CWT-processed ECG (as used in our study) has shown 80% sensitivity and 84% specificity for abnormal relaxation, assessed by low e' (AUC 0.9) in a cohort of patients presenting with symptoms of CAD (5). As in our study, a machine learning approach (random forest classifier) was used but with far more features (n=257), owing to different methodologies. Furthermore, in a larger patient cohort with similar characteristics (e.g. suspected CAD or indication for LV function evaluation) (n=1202), a machine learning algorithm was trained with ewECG features to quantitatively predict e' (6). This algorithm was able to discriminate guideline-defined thresholds with an AUC of 0.84, and given the model generated a continuous output for e', inaccuracies associated with age-based declines could be avoided. While we chose cut-offs, there is no suggestion that in normal aging GLS declines (27), and our definition of diastolic dysfunction would be inclusive of signs of early disease prior to elevations in left atrial pressure. We also believe

that our study population is the most appropriate choice for testing and application of this technology i.e. where echocardiography may not be strictly indicated.

The benefit of our machine-learning method, as opposed to an AI approach (e.g. neural networks), is the potential for interrogation of the model to provide mechanistic insight. We were interested to see whether systolic dysfunction was exclusively, temporally associated with depolarization features, which it was not. This may not be surprising for two reasons: 1) the surface ECG is a simplification of electrical activity spreading across the complex 3D structure of the heart and body and 2) early LV systolic dysfunction and diastolic dysfunction often coexist (28,29). The predictive model for diastolic abnormalities included measures from depolarization as well as repolarization, and most of the features within the low GLS and diastolic models also appeared in the composite LVD model. Clearly, investigation concerning the association between LV dysfunction and specific CWT signatures is in its infancy and is likely to be facilitated by machine-learning algorithms.

Screening for LV dysfunction. The detection of subclinical LVD fulfills some but not all criteria for screening (30). On an individual and population health level, HF is burdensome, and its natural history involves an early asymptomatic stage that is readily detected by abnormal GLS and DD, which carry risk of symptomatic HF and mortality (31-34), analogous to standard markers of impaired LV function (34,35). In terms of treatment, guideline-advocated therapies (ACE-I, ARBs and beta-blockers) significantly improve outcomes (36,37), not only in populations with ischemic cardiomyopathy with reduced ejection fraction (EF), but also in subjects with reduced GLS and diastolic abnormalities with preserved EF, where intensification of cardioprotective therapies may reduce progression to symptomatic HF (38).

Although echocardiography is safe and accurate, cost and access may be problematic. In this setting, the high sensitivity of the ewECG cutoffs that we have developed minimizes

the number of patients going to echocardiography, while at the same time minimizing the numbers of false negatives who do not proceed. The test is low risk and acceptable to patients. The sensitivity of ewECG in our study is superior to the fecal occult blood test for colorectal cancer screening, although specificity is lower (39). However, the risks associated with further testing after a positive ewECG (i.e. echocardiography) are far lower than for colonoscopy, for example. Nonetheless, further work with ewECG will need to include integration of machine-learning algorithms into the device's software to enable immediate interpretation and guide decision-making, and integration of ewECG into clinical workflows.

The alternative is the use of natriuretic peptides (NP) (e.g. BNP ≥50 pg/ml) to guide therapy. Intensification of RAAS and beta-blockade in diabetics with NT-proBNP >125 pg/ml has been shown to reduce cardiovascular (CV) hospitalizations compared with usual care (RR 0.52, 95% CI 0.4-0.68) (40). However, although previous work has shown NP-based therapy reduces asymptomatic LVD in individuals >40 years with CV risk factors (adjusted OR 0.6, 95% CI 0.39-0.93), that study showed no significant difference in HF hospitalization over the 4.2 year mean follow-up (41). Indeed, we found that NT-proBNP had poor screening performance. An inherent problem of BNP in this setting is that levels are artefactually reduced in the setting of obesity. Thus, the role of NT-proBNP in a screening role in this population remains unclear.

Limitations. Machine-learning models are inherently limited by the amount of data available to train the algorithm. Continued acquisition of ewECG data will continue to improve our models. We demonstrated that models differ between targets; performance for one cardiac abnormality in one population should not be extrapolated to others. The poor performance of ewECG for diastolic and early systolic dysfunction abnormalities observed for are likely due to the small number of abnormal studies available for the algorithm to train; as well as the fact that the over-represented group (in this case normal studies) is, by chance, more likely to

be predicted. More developmental work is needed to apply ewECG in these settings. Our study is cross-sectional and therefore we do not know what proportion will go on to develop symptomatic heart failure, or whether ewECG varies between those who do or do not progress. Furthermore, it is unknown whether ewECG can reveal abnormalities before the onset of early LVD or in the subset of patients who fail to exhibit resting echocardiographic abnormalities prior to manifesting HF.

The definition of diastolic dysfunction used in this paper is not conventional. We chose not to use the standard criteria because many subjects are identified as indeterminate. To create a definition suitable for screening, we used clear criteria of raised LA pressure (E/e' with LAE), or if ambiguous (eg isolated LAE), partnered that with another diastolic dysfunction marker. Hence there were three criteria: E/e' >15, E/e'>10 with LAE or impaired relaxation and LAE.

Conclusion. Patients with subclinical LV dysfunction are at increased risk of HF, which may be prevented by initiation of cardioprotective therapy. However, there is currently no consensus as to whether (or how) subclinical LV dysfunction should be detected. Advanced analysis of a routinely acquired ECG using CWT signal processing and machine learning would be a suitably sensitive first step in a selective echocardiographic screening process for detection of LVD. Should such screening be adopted, it could reduce the number undergoing echocardiography by over half.

Perspectives

Clinical Competency (Systems-based practice). The detection of LV dysfunction by echocardiography in asymptomatic people with heart failure (HF) risk factors identifies a group who are at increased risk of HF. However, echocardiographic screening of the population provides logistic and financial challenges. A selection process for echocardiography would make detection more feasible.

Translational Outlook. This study provides data to support the use of "energy waveform" (ew) ECG to identify people at low risk. This could reduce the need for echocardiography by >50%, while at the same time missing a minimal number of patients with LVD. Further evaluation of ewECG is warranted for selection of patients for screening for LVD.

References

1. Hunt SA, Abraham WT, Chin MH et al. 2009 Focused Update Incorporated Into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults. Circulation 2009;119:e391-e479.
2. Garbi M, Edvardsen T, Bax J et al. EACVI appropriateness criteria for the use of cardiovascular imaging in heart failure derived from European National Imaging Societies voting. Eur Heart J Cardiovasc Img 2016;17:711-721.
3. Crowe J, Gibson N, Woolfson M, Somekh M. Wavelet transform as a potential tool for ECG analysis and compression. J Biomed Eng 1992;14:268-272.
4. Meste O, Rix H, Caminal P, Thakor N. Ventricular late potentials characterization in time-frequency domain by means of a wavelet transform. IEEE Trans Biomedi Eng 1994;41:625-634.
5. Sengupta P, Kulkarni H, Narula J. Prediction of Abnormal Myocardial Relaxation From Signal Processed Surface ECG. J Am Coll Cardiol 2018;71:1650.
6. Kagiyama N, Piccirilli M, Yanamala N et al. Machine Learning Assessment of Left Ventricular Diastolic Function Based on Electrocardiographic Features. J Am Coll Cardiol 2020;76:930.
7. Sauer A, Wilcox J, Andrei A, Passman R, Goldberger J, Shah S. Diastolic Electromechanical Coupling. Circ Arrhythm Electrophysiol 2012;5:537-543.
8. Yang H, Wang Y, Nolan M, Negishi K, Okin P, Marwick T. Community Screening for Nonischemic Cardiomyopathy in Asymptomatic Subjects > 65Years With Stage B Heart Failure. Am J Cardiol 2016;117:1959-1965.
9. Willems J, Abreu-Lima C, Arnaud P et al. The Diagnostic Performance of Computer Programs for the Interpretation of Electrocardiograms. N Engl J Med 1991;325:1767-1773.

10. Clark E, Sejersten M, Clemmensen P, Macfarlane P. Automated Electrocardiogram Interpretation Programs Versus Cardiologists' Triage Decision Making Based on Teletransmitted Data in Patients With Suspected Acute Coronary Syndrome. Am J Cardiol 2010;106:1696-1702.
11. The MathWorks I. Five Easy Steps to a Continuous Wavelet Transform. Retrived from: http://matlabizmiranru/help/toolbox/wavelet/ch01_i15html 1994-2005.
12. Potter E, Marwick T. Detection of Stage B Heart Failure in the Community using Energy Waveform ECG. J Am Coll Cardiol 2019;73:Supplement 1.
13. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in python. J Machine Learn Res 2011;12:2825-2830.
14. Pencina MJ, D'Agostino RB, Sr., Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Statistics in medicine 2012;31:101-113.
15. Pencina MJ, D' Agostino Sr RB, D' Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008;27:157-172.
16. Shah A, Claggett B, Loehr L, Chang P, Matsushita K, Kitzman D. Heart Failure Stages Among Older Adults in the Community. Circulation 2017;135:224-240.
17. Administration on Aging (AoA) AfCL, U.S. Department of Health and Human Services. A Profile of Older Americans. 2017.
18. Attia Z, Kapa S, Lopez-Jimenez F et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Med 2019;25:70-74.

19. Schlegel T, Kulecz W, Feiveson A, Greco E, DePalma J, Starc V. Accuracy of advanced versus strictly conventional 12-lead ECG for detection and screening of coronary artery disease, left ventricular hypertrophy and left ventricular systolic dysfunction. BMC Cardiovasc Disord 2010;10:28-28.

20. Addison P. Wavelet transforms and the ECG: a review. Physiological Measurement 2005;26:R155.

21. Reinhardt L, Mäkijärvi M, Fetsch T, Montonen J, Sierra G. Predictive value of wavelet correlation functions of signal-averaged electrocardiogram in patients after anterior versus inferior myocardial infarction. J Am Coll Cardiol 1996;27:53-59.

22. Chevalier P, Rodriguez C, Bontemps L et al. Non-invasive testing of acquired long QT syndrome: evidence for multiple arrhythmogenic substrates. Cardiovasc Res 2001;50:386-98.

23. Savoye C, Klug D, Denjoy I et al. Tissue Doppler echocardiography in patients with long QT syndrome. Eur J Echocardiog 2003;4:209-213.

24. Haugaa K, Edvardsen T, Leren T et al. Left ventricular mechanical dispersion by tissue Doppler imaging: a novel approach for identifying high-risk individuals with long QT syndrome. Eur H J 2009;30:330-337.

25. Piacentino V, Weber C, Chen X et al. Cellular Basis of Abnormal Calcium Transients of Failing Human Ventricular Myocytes. Circ Res 2003;92:651-658.

26. Hobai I, O'Rourke B. Decreased Sarcoplasmic Reticulum Calcium Content Is Responsible for Defective Excitation-Contraction Coupling in Canine Heart Failure. Circulation 2001;103:1577-1584.

27. Potter E, Wright L, Yang H, Marwick T. Normal Range of Global Longitudinal Strain in the Elderly: The Impact of Subclinical Disease. J Am Coll Cardiol Img 2020;https://doi.org/10.1016/j.jcmg.2020.07.014.

28. Yu C-M, Lin H, Yang H, Kong S-L, Zhang Q, Lee S-L. Progression of Systolic Abnormalities in Patients With “Isolated” Diastolic Heart Failure and Diastolic Dysfunction. Circulation 2002;105:1195-1201.

29. Kraigher-Krainer E, Shah A, Gupta D, Santos A, Claggett B, Pieske B. Impaired systolic function by strain imaging in heart failure with preserved ejection fraction. J Am Coll Cardiol 2014;63:447-56.

30. Wilson J, Junger G. Principles and Practice of Screening for Disease. Geneva World Health Organisation Public Health papers No 34 1968.

31. Biering-Sorensen T, Biering-Sorensen S, Olsen F, Sengelov M, Jorgensen P, Mogelvang R. Global Longitudinal Strain by Echocardiography Predicts Long-Term Risk of Cardiovascular Morbidity and Mortality in a Low-Risk General Population: The Copenhagen City Heart Study. Circ Cardiovasc Imaging 2017;10:e005521.

32. Stanton T, Leano R, Marwick T. Prediction of all-cause mortality from global longitudinal speckle strain: comparison with ejection fraction and wall motion scoring. Circ Cardiovasc Imaging 2009;2:356-64.

33. From A, Scott C, Chen H. The Development of Heart Failure in Patients With Diabetes Mellitus and Pre-Clinical Diastolic Dysfunction. J Am Coll Cardiol 2010;55:300-305.

34. Wang Y, Yang H, Huynh Q, Nolan M, Negishi K, Marwick TH. Diagnosis of Nonischemic Stage B Heart Failure in Type 2 Diabetes Mellitus: Optimal Parameters for Prediction of Heart Failure. J Am Coll Cardiol Cardiovasc Img 2018;11:1390-1400.

35. Ammar KA, Jacobsen SJ, Mahoney DW et al. Prevalence and Prognostic Significance of Heart Failure Stages. Circulation 2007;115:1563.

36. The SOLVD Investigators. Effect of Enalapril on Mortality and the Development of Heart Failure in Asymptomatic Patients with Reduced Left Ventricular Ejection Fractions. N Engl J Med 1992;327:685-691.

37. CAPRICORN Investigators. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. Lancet 2001;357:1385-90.

38. Yang H, Negishi K, Wang Y, Nolan M, Marwick T. Imaging-Guided Cardioprotective Treatment in a Community Elderly Population of Stage B Heart Failure. J Am Coll Cardiol Imag 2017;10:217-226.

39. Park D, Ryu S, Kim Y. Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at average risk undergoing colorectal cancer screening. Am J Gastroenterol 2010;105:2017-25.

40. Sweeney C, Ryan F, Ledwidge M et al. Natriuretic peptide-guided treatment for the prevention of cardiovascular events in patients without heart failure. Cochrane Database of Systematic Reviews 2019.

41. Ledwidge M, Gallagher J, Conlon C et al. Natriuretic peptide-based screening and collaborative care for heart failure: the STOP-HF randomized trial. JAMA 2013;310:66-74.

Figure legends

Figure 1: **Conventional ECG traces and corresponding ewECG scalograms after signal processing using CWT**. a) a normal ewECG [determined by MyoVista proprietary software and our machine-learning algorithm]. Echocardiogram was normal. b) Predicted abnormal by our machine-learning algorithm. Participant had abnormal systolic function (GLS 15%), of potential significance is the lower energy associated with the QRS and c) Predicted abnormal by our machine-learning algorithm. Participant had diastolic dysfunction, note low energy associated with the T wave.

Figure 2: Performance of the random forest classifier model utilizing energy waveform ECG features for prediction of LV dysfunction. The area under the receiver operator characteristic (ROC) curves were similar for ewECG features combined with the ARIC-HF risk score (blue AUC 0.83, sensitivity 85% specificity 72%) and with ewECG alone (red AUC 0.78, sensitivity 88% specificity 70%).

Central illustration: Comparison of strategies for LVD screening. The use of an echo screening strategy in all people with risk factors would identify almost 50% of the population as having LVD. A combination of clinical scoring and BNP would reduce echocardiography by about 1/3rd, and miss few people with LVD. However, the use of ew-ECG would reduce the need for echocardiography and miss even fewer patients with LVD.

Table 1: Baseline characteristics by subclinical heart failure stage.

	Stage A HF (n=227)	**LV dysfunction (n=171)**	**p-value**
Clinical and biomarkers			
Age, yrs. (IQR)	68 (62-71)	71 (68-75)	<0.001
Gender (% female)	137 (60)	90 (40)	0.08
Hypertension (%)	185 (82)	152 (90)	0.04
Type II Diabetes Mellitus (%)	41 (18)	60 (35)	<0.001
Atrial fibrillation (%)	9 (4)	13 (8)	0.12
Systolic BP, mmHg (SD)	138 (15)	142 (15)	0.01
Diastolic BP, mmHg (SD)	82 (9)	83 (11)	0.13
Heart rate, beats per min (SD)	63 (9)	66 (10)	0.002
BMI, g/m^2 (SD)	31 (5)	32 (6)	0.09
ACE-I/ARB* (%)	118 (75)	118 (73)	0.64
Beta-Blockers* (%)	17 (11)	25 (15)	0.22
NT-proBNP†, pg/ml (IQR)	51 (30-94)	59 (33-101)	0.39
ARIC HF risk score (IQR)	3.6 (1.22-6.6)	7.1 (3.8-12.9)	<0.001
Standard ECG abnormalities			
Atrial fibrillation (%)	1 (0.4)	4 (2.3)	0.09
LBBB (%)	0 (0)	3 (1.6)	0.05
LV hypertrophy (%)	7 (3)	10 (6)	0.18
Abnormal ECG (per Glasgow analysis) (%)	35 (15)	62 (36)	<0.001
Echocardiographic measures			
LV mass index, g/m^2 (SD)	67 (16)	71 (22)	0.01
LV ejection fraction, % (SD)	64 (6)	61 (7)	<0.001
Global longitudinal strain, % (IQR)	20 (18.9-21)	17 (15.4-18.6)	<0.001
E/A ratio (SD)	0.95 (0.28)	0.80 (0.24)	<0.001
Average e', cm/s (SD)	8.1 (1.7)	7.1 (1.9)	<0.001

Average E/e' (IQR)	8.3 (7.2-9.8)	9.3 (7.3-11.7)	0.003
LAVI, ml/m^2 (IQR)	30 (25-34)	37 (29-42)	<0.001

HF – heart failure, LV – left ventricular, BP – blood pressure, BMI – body mass index, ACE-I/ARB – angiotensin converting enzyme inhibitor/receptor blocker, NT-proBNP – N terminal pro B-type natriuretic peptide, ARIC – Atherosclerosis Risk In Communities, LBBB – left bundle branch block, LAVI – left atrial volume indexed to body surface area. *not available in the Canberra group. †available only in the training dataset.

Table 2: Baseline and Outcome characteristics by training versus test dataset.

	Training (n=287)	**Test (n=111)**	**p-value**
Age, yrs. (IQR)	71 (68-74)	61 (59-66)	<0.001
Gender (% female)	171 (60)	54 (49)	0.05
Hypertension (%)	252 (88)	85 (77)	0.005
Type II Diabetes Mellitus (%)	92 (32)	9 (8)	<0.001
Systolic BP, mmHg (SD)	142 (14)	134 (16)	<0.001
Diastolic BP, mmHg (SD)	84 (9)	79 (10)	<0.001
Heart rate, beats per min (SD)	65 (9)	64 (10)	0.48
BMI, g/m^2 (SD)	32 (6)	31 (5)	0.23
ARIC HF risk score (IQR)	6.3 (3.8-10.6)	1.2 (0.8-2.6)	<0.001
Standard ECG abnormalities			
Atrial fibrillation (%)	5 (1.7)	0 (0)	0.16
LBBB (%)	3 (1)	0 (0)	0.28
LV hypertrophy (%)	15 (5)	2 (1.8)	0.13
Abnormal ECG (per Glasgow analysis) (%)	81 (28)	16 (14)	0.004
Echocardiographic measures			
LV mass index, g/m^2 (SD)	68 (20)	69 (17)	0.53
LV hypertrophy (%)	17 (6)	4 (4)	0.35
LV ejection fraction, % (SD)	63 (7)	62 (5)	0.65
Global longitudinal strain, % (IQR)	19 (17-20)	20 (18-21)	<0.001
GLS ≤ 16% (%)	54 (19)	7 (6)	0.002
E/A ratio (SD)	0.83 (0.22)	1.03 (0.33)	<0.001
Average e', cm/s (SD)	7.8 (1.9)	7.4 (1.7)	0.03
Average E/e' (IQR)	8 (7-10)	9 (8-11)	0.001
LAVI, ml/m^2 (IQR)	35 (30-41)	25 (22-31)	<0.001
Diastolic abnormality (%)	77 (27)	14 (13)	0.002
LV dysfunction (%)	146 (51)	25 (23)	<0.001

LV – left ventricular, BP – blood pressure, BMI – body mass index, ARIC – Atherosclerosis Risk In Communities, HF – heart failure, LBBB – left bundle branch block, LAVI – left atrial volume indexed to body surface area

Table 3: Model features and relative importance (as a proportion of 1) for predicting LV dysfunction.

CWT feature description	ECG Lead(s)	Variable importance
		0.004
Repolarization late measure (RV) (programmed cut-off)		0.01
Repolarization early measure (RV) (programmed cut-off)		0.01
Repolarization late measure (programmed ratio)		0.002
Depolarization measure averaged from Q, R and S waves	aVF	0.1
Minimum energy in early repolarization	V5	0.098
Minimum energy in early repolarization, associated frequency	V5, V6	0.056, 0.055
Minimum energy at peak repolarization	II	0.095
Maximum energy in early repolarization, associated frequency	V6	0.09
Minimum energy in late repolarization, associated frequency	V4, II	0.051, 0.065
Repolarization late measure, associated frequency	I, aVF	0.086, 0.092
Relative deflection of R and S waves	aVR	0.012
Power spectrum (harmonic) ratio feature (LOA sign)	II, V1	0.008, 0.01
Power spectrum (harmonic) ratio feature (HIA sign)	II, III, V4, aVR	0.005, 0.006, 0.005, 0.003
Power spectrum (harmonic) ratio feature (P43 sign)	aVF, V5	0.018, 0.015
Power spectrum (harmonic) ratio feature (P51 sign)	II, III, V1, V3	0.013, 0.014,

		0.017, 0.013
Power spectrum (harmonic) ratio feature (P53 sign)	V4, V6	0.011, 0.009
Power spectrum (harmonic) ratio feature (LO1 sign)	V2, V3	0.007, 0.012
Power spectrum (harmonic) ratio feature (LO3 sign)	I, aVL	0.011, 0.015

LV – left ventricular, CWT – continuous wave transform, ECG – electrocardiogram, RV – right ventricular

Table 4: Performance of random forest classifier models for predicting LV dysfunction, low global longitudinal strain (GLS) alone and diastolic abnormalities.

Prediction target: LV dysfunction **Model components: ewECG features + ARIC HF risk score**		
	Training (n=287)	**Test (n=111)**
Sensitivity	67%	85%
Specificity	68%	72%
ROC AUC (95% CI)	0.71 (0.64-0.77)	0.83 (0.74-0.92)
F-score	0.68	0.60
Prediction target: LV dysfunction **Model components: ewECG features**		
	Training	**Test**
Sensitivity	66%	88%
Specificity	60%	70%
ROC AUC (95% CI)	0.66 (0.59-0.72)	0.78 (0.67-0.88)
F-score	0.64	0.59
Prediction target: Low GLS		
	Training	**Test**
Sensitivity	35%	57%
Specificity	95%	90%
ROC AUC (95% CI)	0.67 (0.58-0.76)	0.65 (0.37-0.93)
F-score	0.45	0.36
Prediction target: Diastolic abnormalities		
	Training	**Test**
Sensitivity	36%	50%
Specificity	94%	90%
ROC AUC 95% (CI)	0.69 (0.62-0.76)	0.62 (0.42-0.82)
F-score	0.47	0.45

LV – left ventricular, ewECG – energy waveform electrocardiogram, ARIC – Atherosclerosis Risk in Communities, HF – heart failure, ROC – receiver operating characteristic curve, AUC – area under curve.

Appendix F

Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource

Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource

The emergence of the COVID-19 pandemic has spurred a global rush to uncover basic biological mechanisms to inform effective vaccine and drug development. Despite the novelty of the virus, global sequencing efforts have already identified genomic variation across isolates. To enable easy exploration and spatial visualization of the potential implications of SARS-CoV-2 mutations in infection, host immunity and drug development, we have developed COVID-3D (http://biosig.unimelb.edu.au/covid3d/).

Stephanie Portelli, Moshe Olshansky, Carlos H. M. Rodrigues, Elston N. D'Souza, Yoochan Myung, Michael Silk, Azadeh Alavi, Douglas E. V. Pires and David B. Ascher

Declared a global pandemic on 11 March 2020, COVID-19 has become the most recent modern-day global health challenge, infecting 10 million people and claiming more than 500,000 lives within 6 months of being reported to the World Health Organization. Consequently, the scale of its humanitarian and economic impact has driven academic and pharmaceutical efforts to develop vaccines and antiviral treatments. Current efforts include more than 118 active vaccine candidates and numerous additional endeavors to identify biologics and small-molecule treatments.

One further challenge in controlling COVID-19 is the accumulation of variation across genes. Sources indicate that SARS-CoV-2 is mutating at approximately two variants per month, but the potential effects of the accumulation of these variants (Supplementary Fig. 1) on molecular diagnostics and the development of candidate vaccines and treatments remain poorly explored. Fortunately, the continual rapid increase in the amount of SARS-CoV-2 genome sequence data and structural information available provides an opportunity to analyze both data sources concomitantly, thus presenting a unique opportunity to not only understand how variants might affect patient outcomes, but also anticipate and minimize their potential roles in viral escape through early incorporation of this information within the development pipeline.

To facilitate such an understanding, we have developed a comprehensive online resource, COVID-3D, to enable analysis and interpretation of more than 11,000 variants detected in circulating SARS-CoV-2 genomic sequences (Supplementary Fig. 2). We have mapped these circulating variants and their frequencies to the corresponding protein sequences (Supplementary Table 1) and structures of the SARS-CoV-2 proteins derived from available experimental information (Supplementary Table 2), thus permitting direct comparison of variant clustering between the sequence and structural representations, along with the identification of coevolutionary relationships and potential compensatory mutations. Beyond these circulating variants, we have identified mutations from the longer-circulating related viruses BAT RaTG13 and SARS-CoV, to enable further investigation of the mutations that drove the species jump from RaTG13 and that increased the infectivity and mortality beyond those of SARS-CoV. Our interactive three-dimensional viewer enables fast and intuitive spatial visualization of SARS-CoV-2 variants, highlighting their potential effects on protein structure and interactions[1–7] (Supplementary Figs. 3–6). This viewer is particularly useful for analyzing sites that are currently being targeted by potential therapeutics. A built-in mutation-analysis tool allows users to contrast properties and identify patterns in the data, plotting correlations and distributions (Supplementary Fig. 7).

To further enhance therapeutic discovery efforts, we have included maps of the fragment-binding hotspots to capture likely drug-binding sites[8,9], as well as predicted antigenicity maps[10,11] on the structures, which permit rational selection of target sites and compound design, specifically avoiding already circulating variants (Supplementary Fig. 4). Finally, combining this structural information with evolutionary and population variation analysis can further aid in identifying sites that are relatively less likely to accommodate mutations in the future. To facilitate this analysis, COVID-3D also allows users to go from analyzing a protein pocket to virtual screening in several clicks[12]. In an illustrative example, we have used COVID-3D to provide insights into the two main therapeutic targets: the spike protein and main proteinase.

The SARS-CoV-2 spike protein binds human angiotensin-converting enzyme 2 (ACE2), which mediates cell entry. Subsequently, the spike protein's ACE2-receptor-binding domain has been the main target of most vaccine programs. Measures of selective pressure suggest that the spike protein is one of the viral proteins most tolerant to the introduction of mutations[13,14] (Supplementary Table 1). Closer inspection (http://biosig.unimelb.edu.au/covid3d/protein/QHD43416/CLOSED) indicates that although SARS-CoV-2 was discovered only 6 months before the time of analysis, substantial variation can already be seen across the protein surface, including in predicted epitope regions in the receptor-binding domain (Fig. 1). Of these variants, QHD43416 p.Asp614Gly is present in two-thirds of the sequenced strains, although its actual importance remains unclear, despite initial suggestions that it may increase transmissibility[15]. The residue is located far from the ACE2 interface (73 Å) and has been predicted to have a mildly stabilizing effect on protein stability (0.5 kcal mol^{-1} according to DUET[3] and 2.3 kcal mol^{-1} according to SDM[2] analyses) and hence a minimal fitness cost[16]. However, it has been predicted to alter protein dynamics and the interactions between the subunits (4.4 Å from the interface; −0.5 kcal mol^{-1} for

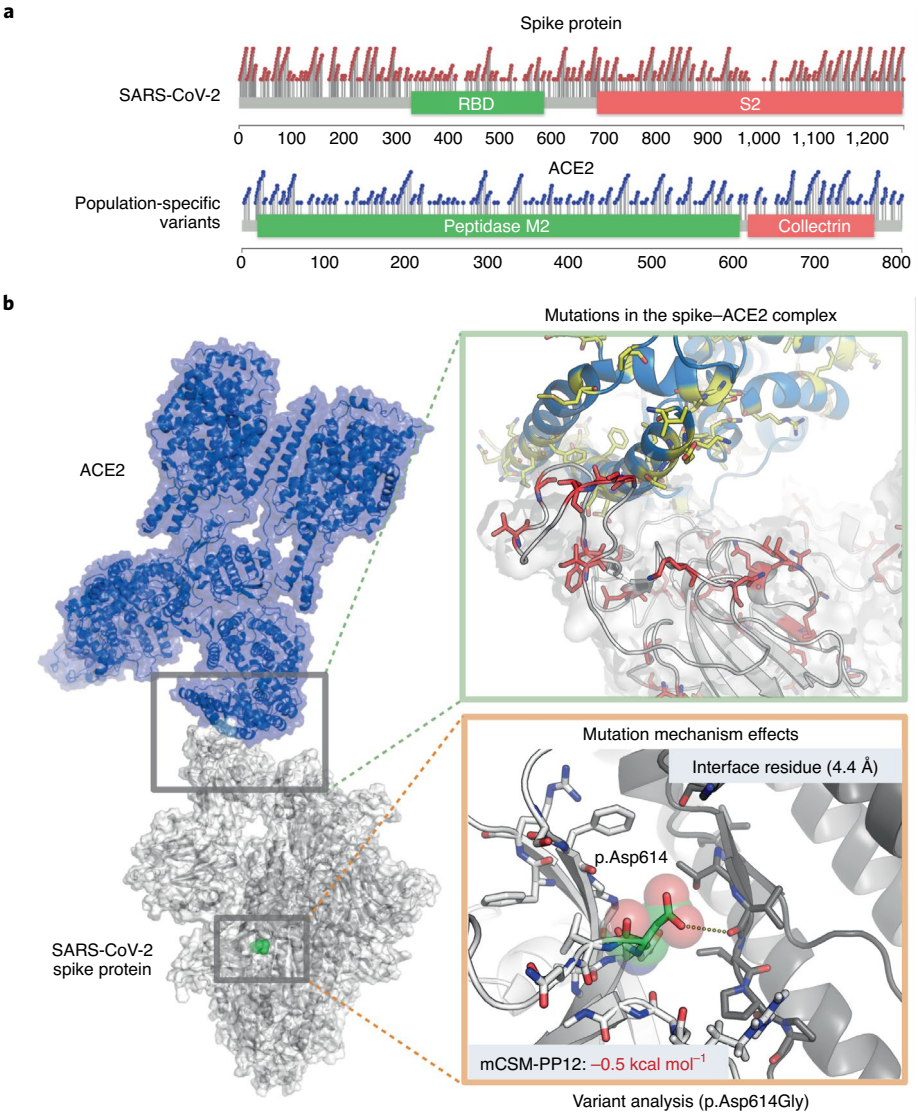


Fig. 1 | Population variation across the spike–ACE2 complex. a, Lollipop plots of circulating missense variants in the SARS-CoV-2 spike protein and population-specific missense variants in human ACE2 illustrate the broad distribution of variants across the proteins. **b**, When visualized spatially, several variants seen at the ACE2–spike interface are predicted to affect the binding affinity. One of the most prevalent circulating SARS-CoV2 spike variants, p.Asp614Gly, is located far from the ACE2 interface but close to the spike-trimer interface and is predicted to lead to structural perturbations.

the closed form versus –0.35 kcal mol^{-1} for the open form, according to mCSM-PPI2 analysis[6]), thus potentially affecting the equilibrium between open and closed states.

Interestingly, when we examined population-specific variants across ACE2, we observed several population-specific variants across the interface recognized by the spike protein (Fig. 1a). Evaluation of the consequences of these variants with mCSM-PPI[6], which has been experimentally validated on this protein system[17], shows potential significant effects on the binding affinity of spike protein, thus paving the way for further work exploring the influence of these variants on COVID-19 severity and progression.

Apart from the spike protein, the main proteinase (http://biosig.unimelb.edu.au/covid3d/protein/QHD43415_5/APO) has also attracted many therapeutic development efforts as a target for the development of small-molecule inhibitors. The main proteinase, however, is not particularly intolerant to missense variants (Supplementary Table 1), thus potentially promoting the emergence of resistant variants. The structures show that several circulating variants already present in the drug-binding site may have effects on efficacy (Fig. 2a). Using COVID-3D, we leveraged the abundance of SARS-CoV-2 genomic sequences to calculate measures of mutational tolerance, and we identified several genes under strong purifying selection (Supplementary Table 1). These include the genes encoding helicase, RNA polymerase, NSP4, NSP9 and ExoN, which may serve as novel, promising drug targets with few circulating variants seen near the druggable pockets (Fig. 2b).

COVID-3D provides an easy-to-use bridge between genomic information and structural insights to better guide biological understanding and treatment efforts. The data and code (http://biosig.unimelb.edu.au/covid3d/code) are freely available via the web interface (http://biosig.unimelb.edu.au/covid3d/). As new structural and sequence data become available, COVID-3D will be periodically updated to enable

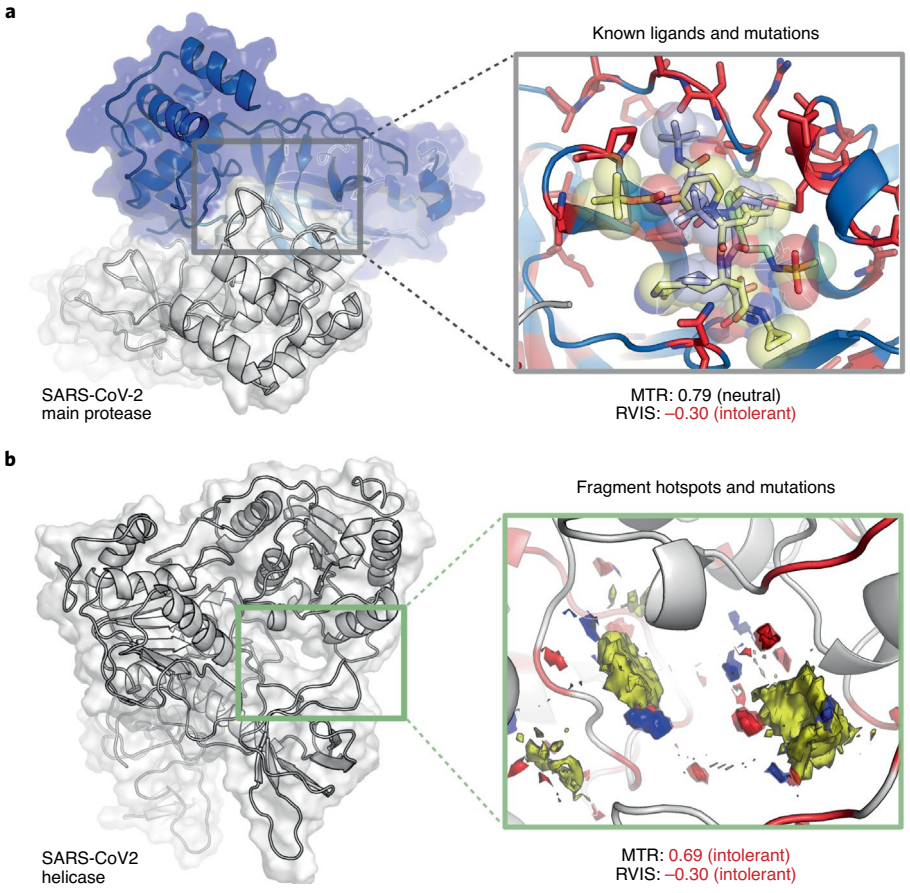


Fig. 2 | Visualization of SARS-CoV-2 circulating variants relative to druggable pockets. a, The gene encoding the main proteinase is neutral to the introduction of missense variants, with an overall missense tolerance score (MTR) and residual variation intolerance score (RVIS) both indicating that the gene is tolerant to genetic variation. Some circulating variants (red sticks) have already been observed to lead to alterations near binding sites of known inhibitors (boceprevir shown in yellow) and are likely to affect drug binding. Therefore, resistance mutations could be selected for with widespread use. **b**, The gene encoding helicase is among the SARS-CoV-2 genes most intolerant to missense variation, with low MTR and RVIS scores. Mapping the fragment-binding hotspots of the protein shows pockets with apolar (yellow), hydrogen-bond-donor (blue) and hydrogen-bond-acceptor (red) potential. Although some variation has been observed near this region, optimization of interactions to avoid these sites could decrease the potential for future resistance.

their integration into ongoing efforts to understand and combat SARS-CoV-2. ❐

Stephanie Portelli[1,2,5], Moshe Olshansky[1,2,5], Carlos H. M. Rodrigues[1,2,5], Elston N. D'Souza[1,2,5], Yoochan Myung[1,2], Michael Silk[1,2], Azadeh Alavi[1,2], Douglas E. V. Pires[1,2,3,5] and David B. Ascher[1,2,4]✉

[1]*Structural Biology and Bioinformatics, Department of Biochemistry, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia.* [2]*Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia.* [3]*School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia.* [4]*Department of Biochemistry, University of Cambridge, Cambridge, UK.* [5]*These authors contributed equally: Stephanie Portelli, Moshe Olshansky, Carlos H. M. Rodrigues, Elston N. D'Souza, Douglas E. V. Pires.*
✉*e-mail: david.ascher@unimelb.edu.au*

Published online: 9 September 2020
https://doi.org/10.1038/s41588-020-0693-3

References

1. Jubb, H. C. et al. *J. Mol. Biol.* **429**, 365–371 (2017).
2. Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B. & Blundell, T. L. *Nucleic Acids Res.* **45**, W229–W235 (2017).
3. Pires, D. E., Ascher, D. B. & Blundell, T. L. *Nucleic Acids Res.* **42**, W31–W319 (2014).
4. Pires, D. E., Ascher, D. B. & Blundell, T. L. *Bioinformatics* **30**, 335–342 (2014).
5. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. *Nucleic Acids Res.* **46**, W350–W355 (2018).
6. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. *Nucleic Acids Res.* **47**, W338–W344 (2019).
7. Pires, D. E., Blundell, T. L. & Ascher, D. B. *Sci. Rep.* **6**, 29575 (2016).
8. Radoux, C. J., Olsson, T. S. G., Pitt, W. R., Groom, C. R. & Blundell, T. L. *J. Med. Chem.* **59**, 4314–4325 (2016).
9. Kawabata, T. *Proteins* **78**, 1195–1211 (2010).
10. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. *Nucleic Acids Res.* **45**, W24–W29 (2017).
11. Ponomarenko, J. et al. *BMC Bioinforma.* **9**, 514 (2008).
12. Pires, D. E. V. et al. *Bioinformatics* **36**, 4200–4202 (2020).
13. Silk, M., Petrovski, S. & Ascher, D. B. *Nucleic Acids Res.* **47**, W121–W126 (2019).
14. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. *PLoS Genet.* **9**, e1003709 (2013).
15. Korber, B. et al. *Cell* **182**, 812–827.e19 (2020).
16. Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G. & Furnham, N. *Sci. Rep.* **8**, 15356 (2018).
17. MacGowan, S. A. & Barton, G. J. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.03.074781 (2020).

Acknowledgements

We thank N. Thanh Binh (Bioinformatics Institute, A*STAR, Singapore) for help with molecular dynamics simulations. S.P., C.H.M.R. and Y.M. were supported by a Melbourne Research Scholarship. D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG; MR/M026302/1) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). D.B.A. was funded by the Jack Brockhoff Foundation (JBF 4186, 2016); the Wellcome Trust (200814/Z/16/Z) and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). This work was supported in part by the Victorian Government's OIS Program. This research has been conducted using the UK Biobank Resource under Application Number 50000.

Author contributions

S.P. was responsible for structure curation, homology modeling and structural characterization. M.O. was responsible for curating SARS-CoV-2 variants. C.H.M.R. was responsible for developing the website. Y.M. performed the molecular dynamics analysis and assisted with the website. E.N.D. was responsible for curating the human population variants. E.N.D. and M.S. were responsible for calculating intolerance scores. A.A. assisted with SARS-CoV-2 genomic curation. D.E.V.P. was responsible for figures, normal mode analysis and implementation of virtual screening, and assisted with fragment hotspot calculations, website development and supervision. D.B.A. conceived, designed and supervised all aspects of the project and website, and wrote the manuscript. All authors assisted with manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0693-3.

Appendix G

mCSM-membrane: predicting the effects of mutations on transmembrane proteins

Published online 29 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W147–W153***
doi: 10.1093/nar/gkaa416

mCSM-membrane: predicting the effects of mutations on transmembrane proteins

Douglas E.V. Pires[1,2,3,*], Carlos H.M. Rodrigues[1,2] and David B. Ascher[1,2,4,*]

[1]Computational Biology and Clinical Informatics, Baker Institute, Melbourne, Victoria 3004, Australia, [2]Structural Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Parkville, VIC, 3052, Australia, [3]School of Computing and Information Systems, University of Melbourne, Parkville, VIC, 3052, Australia and [4]Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK

Received February 21, 2020; Revised May 04, 2020; Editorial Decision May 05, 2020; Accepted May 28, 2020

ABSTRACT

Significant efforts have been invested into understanding and predicting the molecular consequences of mutations in protein coding regions, however nearly all approaches have been developed using globular, soluble proteins. These methods have been shown to poorly translate to studying the effects of mutations in membrane proteins. To fill this gap, here we report, mCSM-membrane, a user-friendly web server that can be used to analyse the impacts of mutations on membrane protein stability and the likelihood of them being disease associated. mCSM-membrane derives from our well-established mutation modelling approach that uses graph-based signatures to model protein geometry and physicochemical properties for supervised learning. Our stability predictor achieved correlations of up to 0.72 and 0.67 (on cross validation and blind tests, respectively), while our pathogenicity predictor achieved a Matthew's Correlation Coefficient (MCC) of up to 0.77 and 0.73, outperforming previously described methods in both predicting changes in stability and in identifying pathogenic variants. mCSM-membrane will be an invaluable and dedicated resource for investigating the effects of single-point mutations on membrane proteins through a freely available, user friendly web server at http://biosig.unimelb.edu.au/mcsm_membrane.

INTRODUCTION

Integral membrane proteins play an essential role as the gateway to the cell, mediating transport, signalling and adhesion amongst many other functions. Mutations in membrane proteins are associated with a wide variety of common diseases, including heart disease, and consequently have been the site of action for over 50% of small molecule drugs (1). While they represent 20–30% of the genes in the human genome (2–4), they can be challenging to experimentally characterise as they tend to be unstable when extracted from the lipid bilayer. Consequently, less than 0.5% of experimentally determined structures are of integral membrane proteins.

There is therefore an increasing demand for methods capable of identifying mutations that might improve stability, to facilitate structural and functional characterization, and to identify novel disease-causing variants. Increasing computational power offers new opportunities to address these challenges, however most tools have been built using experimental information on predominantly globular, soluble proteins, and that have been shown to poorly translate to predicting the effects of mutations in membrane proteins (5).

The need for methods tailored for investigating mutation effects on transmembrane proteins becomes evident when considering the differences in residue environment in comparison with globular proteins. While many studies involving globular proteins have shown that solvent accessibility and residue depth correlates with mutation effects (6), for example buried and deep residues tend to be more conserved and mutations tend to have larger effects in stability, these might not be applicable for integral membrane proteins. To circumvent this, sophisticated ways to describe and represent residue environments are necessary.

We have previously tackled this task by developing the concept of graph-based signatures and showed they can provide powerful insights into understanding and predicting the effects of mutations on protein structures, including how mutations alter protein stability (6–8), dynamics (8), interactions with other molecules (7–14) and their relation to emergence of genetic diseases (15–27) and drug resistance (10,19,28–38).

Here we introduce mCSM-membrane, a web server that adapts and optimizes our well-established mCSM graph-based signatures framework in order to provide improved

*To whom correspondence should be addressed. David B. Ascher. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au
Correspondence may also be addressed to Douglas E.V. Pires. Email: douglas.pires@unimelb.edu.au

predictive performance of the molecular consequences of mutations in membrane proteins.

MATERIALS AND METHODS

Data sets

The general workflow of mCSM-membrane is shown in Figure 1. mCSM-membrane was trained using two separate data sets of experimentally characterized mutations in transmembrane proteins, for which 3D structures were available.

The first data set contained experimentally measured effects of mutations on protein stability. This was obtained from (5) and encompasses 223 single-point missense mutations on 7 different proteins with experimental crystal structures available in the Protein Data Bank. The mutation effects were obtained in terms of the difference in Gibbs free energy of folding ($\Delta\Delta G = \Delta G_{WT} - \Delta G_{MT}$, in Kcal/mol), with negative values denoting destabilising mutations and positive values denoting stabilising mutations, consistent with previously published methods. As discussed in previous works (8,10,13,14), the original data set was biased towards destabilising mutations (Supplementary Figure S1), which tend to affect machine learning methods. To circumvent this sampling limitation, we have modelled the hypothetic reverse mutations via comparative homology modelling and assigned the same $\Delta\Delta G$ value as the forward mutation, with the opposite signal, in other words: $\Delta\Delta G_{WT\rightarrow MT} = -\Delta\Delta G_{MT\rightarrow WT}$. Only reverse mutations with a measured effect in stability <2 kcal/mol were considered, in order to avoid situations where the reverse mutation could potentially compromise protein folding. Structures for reverse mutations were generated using the mutate function within Modeller (39) followed by refinement. A total of 181 reverse mutations were modelled, leading to a final data set of 404 mutations with associated stability effects (Supplementary Figure S1). Forward and reverse mutations pairs were kept together either in training or test sets. This was further divided into training (342 missense mutations occurring in 4 proteins, PDB IDs 2XOV, 1PY6, 3GP6 and 1QD6; 156 decreasing stability ($\Delta\Delta G < -0.4$ kcal/mol), 56 neutral, 130 increasing stability ($\Delta\Delta G > 0.4$ kcal/mol) and independent blind test (62 mutations occurring in the remaining three proteins, PDB IDs 1QJP, 2K73 and 1AFO, 28 decreasing stability, 14 neutral, 20 increasing stability). Training and test sets used in mCSM-membrane were non-redundant in terms of protein identity (<16% sequence identity – Supplementary Table S1) The proteins were also assessed in terms of their structural similarity using TMAlign and shared no more than 64% similarity.

The second data set was selected in order to train a structure-based model for predicting disease-associated mutations tailored for transmembrane proteins and was collected from (40). It comprises 539 single-point missense mutations in 62 different proteins, labelled either as benign or pathogenic, from the UniProtKB/Swiss-Prot variant database (41) This dataset was also further divided in training set (485 mutations, 347 pathogenic, 138 benign) and independent blind test (54 mutations, 38 pathogenic, 16 benign) for validation purposes, consistent with the data set defined by the BORODA-TM method for comparison purposes. Seven mutations described in the original data set, on two different residues of protein 4ZWJ could not be mapped to the structure available and therefore were removed from the training set. These compose non-redundant datasets, with sequence identity levels less than 50% and less than 75% structural similarity (calculated using TMalign).

The data sets used to develop mCSM-membrane are available to download at http://biosig.unimelb.edu.au/mcsm_membrane/data.

Modelling effects of mutations

Single-point mutations can lead to a range of structural and functional changes. To try to encapsulate and explore the effects of single-point mutations on membrane proteins, we used two classes of structural features, in addition to sequence-based calculations.

Graph-based structural signatures

One of the core components of mCSM-membrane is our well-established approach of using the concept of graph-based structural signatures (mCSM) to represent the environment of the wild-type residue (7) and describe both its geometry and physicochemical properties. Our approach aims to model wild-type residue environments as graphs, where atoms are represented as nodes (labelled based on their properties, i.e. pharmacophores) and their interactions as edges. By varying a distance cut off, different graphs are induced and cumulative distributions of distances for different pharmacophore/interactions generated, composing a concise and effective representation of the residue environment. This information is then used as evidence to train and test predictive methods using supervised learning.

Molecular interactions

To capture information on whether, and how, a single-point mutation disrupted the intricate molecular interaction network, intra-molecular interactions were calculated using Arpeggio (42).

Pharmacophore modelling and sequence-based features

The effect of the mutation on the residue environment is modeled using a pharmacophore representation for residues as previously described (7). Sequence-based features describing protein properties and amino acid composition were also calculated using the BioPython python library (43). These include AAindex amino acid mutation matrices and indexes representing physicochemical properties (44) and ProtParam, for calculating general protein sequence properties, including amino acid composition, molecular weight, isoelectric point, and hydropathicity (45).

Differently from globular proteins, neither residue depth, nor solvent accessibility, showed a significant correlation with stability effects ($r = 0.07$ and $r = 0.09$, respectively. Supplementary Figure S2).

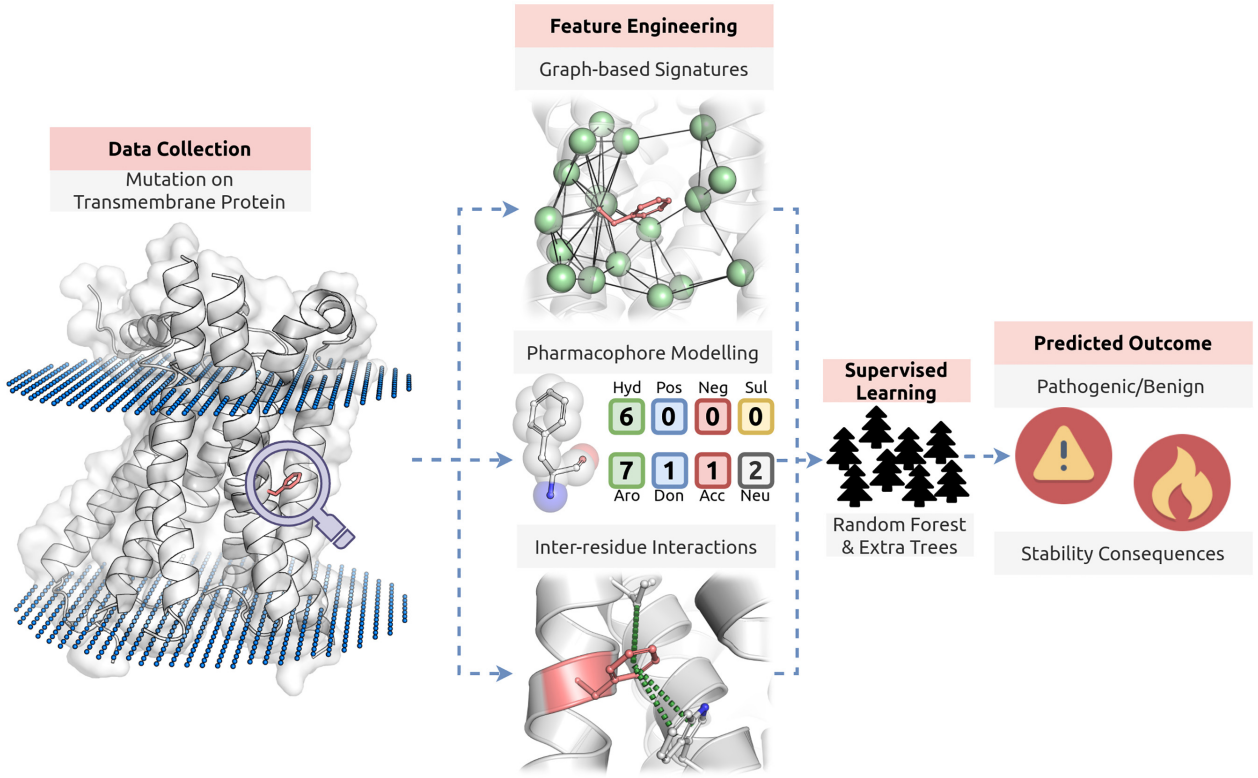


Figure 1. mCSM-membrane workflow. The first methodological step on mCSM-membrane was data collection. Experimentally validated effects of mutations on protein stability and pathogenicity were obtained for transmembrane proteins with available structures. During feature engineering, three main classes of features are generated: (i) graph-based signatures of the wild-type residue environment, (ii) a pharmacophore modelling of mutation effects (together with sequence-based properties) and (iii) the inter-residue interactions established. These are then used as evidence to train and test supervised learning algorithms. Random Forest for classification and Extra Trees for regression were the best performing and, therefore, selected methods.

WEB SERVER

We have implemented mCSM-membrane as a user-friendly and freely available web server (http://biosig.unimelb.edu.au/mcsm_membrane/). The Bootstrap framework version 3.3.7 was used to develop the server front end, while the back-end was built in Python using the Flask framework version 1.0.2. The server is hosted on a Linux server running Apache 2.

Input

mCSM-membrane can be used in two different ways: to either assess the effects of mutations on membrane protein stability, or to assess their pathogenicity (Supplementary Figure S3). For user-specified variations two options are available. The 'Single Mutation' option requires users to provide a PDB file or PDB accession code of the structure of the protein, the point mutation specified as a string containing the wild-type residue one-letter code, its corresponding residue number (consistent with the provided structure) and the mutant residue one-letter code. Alternatively, the 'Mutation List' option allows users to upload a list of mutations in a file for batch processing. For both options, users are also required to specify the chain identifier in which the wild-type residues are located as well as the Uniprot accession code for the protein of interest or provide its sequence in FASTA format. For homo-oligomers, mCSM-membrane will only consider the mutation in the provided chain, however the overall environment (oligomer) will be considered for feature generation.

In order to assist users to submit their jobs for predictions, sample submission entries are available in both submission pages and a help page is also available via the top navigation bar.

Output

For the Stability option, mCSM-membrane outputs the predicted change in membrane protein stability (in kcal/mol), while for the Pathogenicity option mCSM-membrane outputs whether the mutation is predicted as Benign or Pathogenic.

With the Single Mutation option, mCSM-membrane outputs the prediction along with an interactive 3D viewer showing the wildtype residue environment and a depiction of the predicted transmembrane topology using Protter (46) (Supplementary Figure S4). In addition, all non-covalent interactions, generated using Arpeggio, made by the wild-type residue are available for download as a Pymol session file. For the Mutation List option, the results are summarized in a downloadable table from which users can access details for each single variant (Supplementary Figure S5).

VALIDATION

Predicting effects of mutations on transmembrane protein stability

In order to build a robust and reliable model for predicting the effects of mutations on transmembrane stability, mCSM-Membrane was trained using a stratified 10-fold cross-validation approach with 10 bootstrap repetitions. Selection of the blind test was repeated 10 times in a stratified manner, with the model assessed on the remaining data using 10-fold cross-validation, in order to evaluate the robustness of the model. Our method achieved an average Pearson, Spearman and Kendall correlations of 0.72, 0.72 and 0.53, respectively, with a standard deviation of 0.09 across the 10 runs (Figure 2A). We then evaluated the ability of the model to capture destabilizing and stabilizing mutations, using a classification by regression approach. mCSM-Membrane achieved a Mathew's Correlation Coefficient of 0.65 and F1-score of 0.81, correctly capturing 82% of stabilizing and 83% of destabilizing mutations. The effect of considering reverse mutations in the data set was also assessed. When only forward mutations are considered (i.e. removing reverse mutations from training and test sets), performance drops considerably, achieving a Pearson's correlation of 0.58 and a Mathew's Correlation Coefficient of 0.79 and *F*1-score of 0.72, highlighting the importance of considering reverse mutations to balance the data set.

mCSM-Membrane was further evaluated using a blind test set of 62 mutations across 3 proteins, not present in our original training data sets. Our model achieved Pearson, Spearman and Kendall correlations of 0.67, 0.62 and 0.45 (Figure 2B), respectively, consistent with training performance, providing confidence in the generalizability and robustness of our model. Despite the low level of similarity between proteins in training and test sets, and to eliminate any potential selection bias while training and validating our method, we also evaluate the process of selecting an independent test set in a bootstrapped manned 100×, and evaluated the performance of the method on cross validation and test set. mCSM-membrane achieved a correlation of 0.68 (sd = 0.02) on 10-fold cross validation and 0.67 (sd = 0.07) on tests, demonstrating the robustness of the method. Additionally, mCSM-Membrane was compared to well established tools designed to predict the effects of mutations on protein stability. mCSM-Membrane significantly outperformed all tools tested ($P < 0.05$ by Fisher *r*-to-*z* transformation test, Table 1). Consistent with previous results, the other stability predictive tools tested were only weakly predictive across these mutations in transmembrane proteins (Table 1).

Table 1. Comparative performance of mCSM-membrane across training and test data sets with alternative stability predictors

	Training		Test	
Method	Pearson's correlation	RMSE	Pearson's correlation	RMSE
FoldX	0.48*	1.18	0.57	1.25
iMutant	0.27*	1.29	0.37*	1.41
CUPSAT	0.01*	1.34	0.15*	1.50
AUTOMUTE (RepTree)	0.17*	1.32	0.05*	1.52
AUTODMUTE (SVM)	0.14*	1.33	0.04*	1.52
MAESTRO	0.20*	1.16	0.17*	1.09
SDM	0.01*	1.34	−0.14*	1.51
mCSM	0.21*	1.31	0.59	1.23
DUET	0.18*	1.32	0.47*	1.34
Dynamut	0.31*	1.27	0.62	1.19
mCSM-membrane	0.72	0.93	0.67	1.13

*P-value < 0.05 by Fisher *r*-to-*z* transformation test compared to mCSM-membrane

Application to homology models

Experimentally solving structures of transmembrane proteins is particularly challenging. The evolution of comparative homology and threading algorithms, however, has allowed for data augmentation for modelled structures at a proteome-scale (47). To assess the performance of mCSM-membrane on homology models, we have generated models using templates with no more than 37% identity for three different proteins, originally selected as the blind test of our stability predictor. Supplementary Table S2 shows the information on templates used in this process.

Performance on blind test using the homology models deteriorates only slightly ($r = 0.63$. Supplementary Figure S6), compared to performance on experimental structures ($r = 0.68$), highlighting the robustness of the model and ability to accurately predict effects of mutations on homology models. This defines a simple guideline for using mCSM-membrane on homology models.

Identifying pathogenic mutations in transmembrane proteins

The second predictive mode for mCSM-membrane is a predictor capable of accurately distinguishing between pathogenic and benign mutations tailored for transmembrane proteins (Table 2). This predictor was trained and assessed on 10-fold cross validation, with its performance compared to alternative methods available. Our pathogenicity predictor achieved an Mathew's Correlation Coefficient (MCC) of 0.77 and *F*1-score of 0.91 significantly outperforming SIFT (0.43 and 0.85), PolyPhen2 (0.54 and 0.89) PROVEAN (0.48 and 0.85), MutPred2 (0.48 and 0.79), PON-P2 (0.38, 0.71). The only method that achieved a higher performance than mCSM-membrane during cross validation was BORODA-TM (0.87 and 0.96). However, the discrepancy between the reported performance in cross validation and blind test for BORODA-TM (on blind it achieves an MCC of 0.46 and *F*1 of 0.78) is a strong indication of overfitting.

Our predictor was further validated via a blind test achieving an MCC of 0.73 and F1-score of 0.89, performance compatible with cross validation, outperforming alternative methods and demonstrating the efficacy of a transmembrane-specific predictor no identifying pathogenic mutations. Figure 2C and D shows the ROC curves comparing the performance of the four methods during cross validation and blind tests, with our predictor achieving an Area Under the ROC Curve (AUC) of 0.89 and 0.95, respectively.

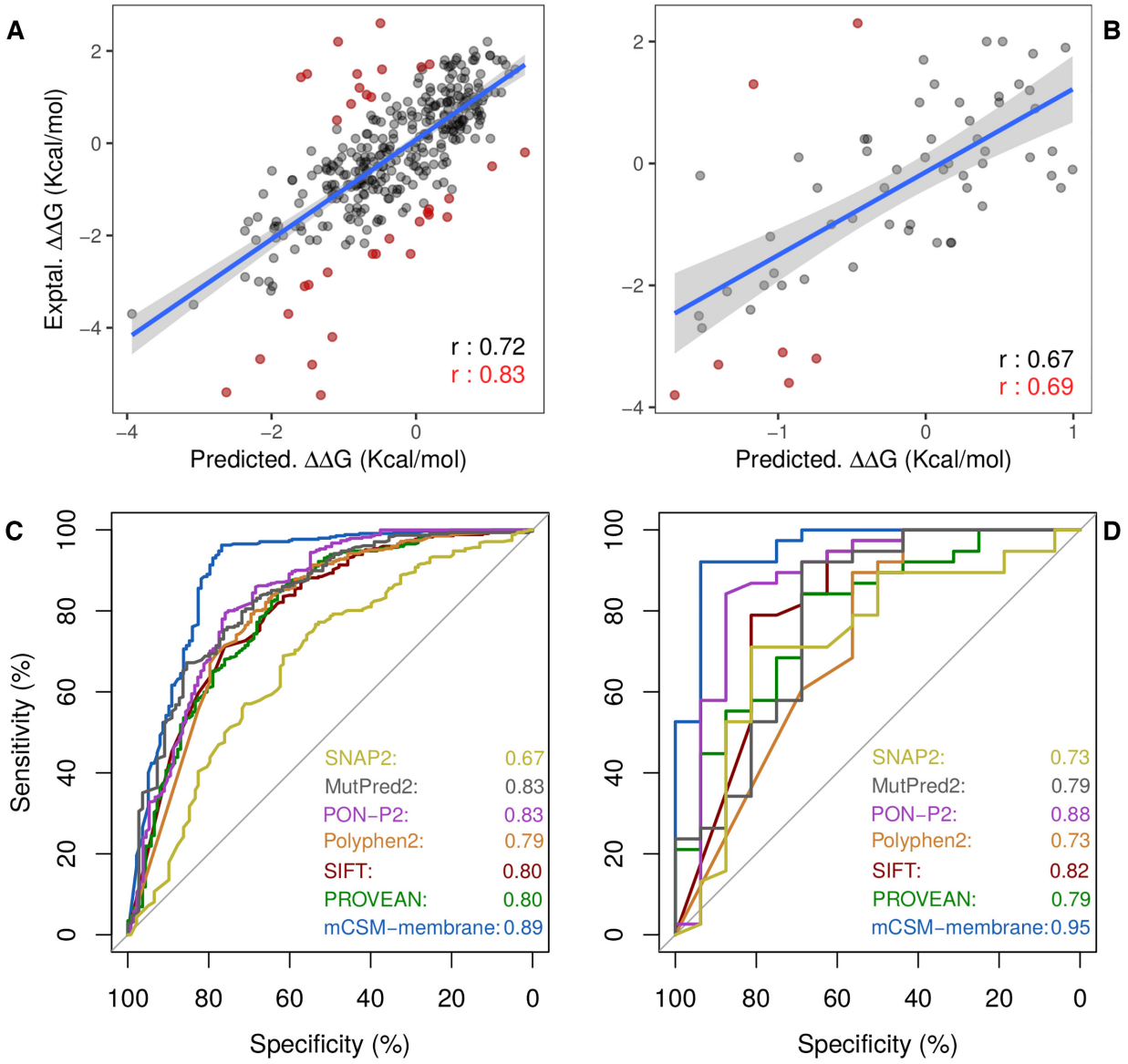


Figure 2. Performance evaluation of mCSM-membrane on cross validation and blind tests. (**A**) shows the performance of mCSM-membrane on predicting effects of mutations on stability for transmembrane proteins during 10-fold cross validation, achieving a Pearson's correlation of 0.72 (0.83 on 90% of the data). During blind test (**B**), mCSM-membrane achieved a correlation of 0.67 with experimental data. For the pathogenicity predictor, (**C**) and (**D**) show the performance of mCSM-membrane in comparison with well-established methods as ROC plots on cross validation and blind test, respectively. Our method achieved AUC of 0.89 and 0.95.

Table 2. Performance assessment of mCSM-membrane in predicting pathogenic mutations across training and test data sets, in comparison with alternative methods.

	Training			Test		
Method	AUC	F1	MCC	AUC	F1	MCC
PolyPhen2	0.79	0.79	0.47	0.73	0.75	0.40
SIFT	0.80	0.77	0.43	0.82	0.84	0.63
PROVEAN	0.80	0.79	0.48	0.79	0.75	0.40
SNAP2	0.67	0.70	0.26	0.73	0.66	0.21
MutPred2	0.75	0.79	0.48	0.75	0.82	0.57
PON-P2	0.83	0.71	0.38	0.88	0.78	0.53
BORODA-TM*	- - -	0.96	0.87	- - -	0.78	0.46
mCSM-membrane	0.89	0.91	0.77	0.95	0.89	0.73

*AUC values were not calculated for BORODA-TM as no scores, rankings or class probabilities were available.

CONCLUSION

Here, we introduce mCSM-membrane, a web server that uses our graph-based signatures to predict the effects of single-point missense mutations on the stability of transmembrane proteins and the likelihood of them being disease associated. The method represents a significant advance upon our current predictive platform, outperforming previous methods, which had been built using globular soluble proteins.

mCSM-membrane is freely available as user-friendly and easy to use web server at http://biosig.unimelb.edu.au/mcsm_membrane/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1]; Jack Brockhoff Foundation [JBF 4186, 2016]; Wellcome Trust [200814/Z/16/Z]; Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405]; Victorian Government's OIS Program (in part). Funding for open access charge: MRC.
Conflict of interest statement. None declared.

REFERENCES

1. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L. (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
2. Frishman,D. and Mewes,H.W. (1997) Protein structural classes in five complete genomes. *Nat. Struct. Biol.*, **4**, 626–628.
3. Fagerberg,L., Jonasson,K., von Heijne,G., Uhlen,M. and Berglund,L. (2010) Prediction of the human membrane proteome. *Proteomics*, **10**, 1141–1149.
4. Babcock,J.J. and Li,M. (2014) Deorphanizing the human transmembrane genome: A landscape of uncharacterized membrane proteins. *Acta Pharmacol. Sin.*, **35**, 11–23.
5. Kroncke,B.M., Duran,A.M., Mendenhall,J.L., Meiler,J., Blume,J.D. and Sanders,C.R. (2016) Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. *Biochemistry*, **55**, 5002–5009.
6. Pandurangan,A.P., Ascher,D.B., Thomas,S.E. and Blundell,T.L. (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.*, **45**, 303–311.
7. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–W319.
8. Rodrigues,C.H., Ascher,D.B. and Pires,D.E. (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. *Nucleic Acids Res*, **46**, W127–W132.
9. Pires,D.E. and Ascher,D.B. (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res*, **44**, W469–W473.
10. Phelan,J., Coll,F., McNerney,R., Ascher,D.B., Pires,D.E., Furnham,N., Coeck,N., Hill-Cawthorne,G.A., Nair,M.B., Mallard,K. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med*, **14**, 31.
11. Pires,D.E. and Ascher,D.B. (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res.*, **44**, W557–W561.
12. Pires,D.E.V. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.
13. Myung,Y., Rodrigues,C.H.M., Ascher,D.B. and Pires,D.E.V. (2020) mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics*, **36**, 1453–1459.
14. Pires,D.E.V., Rodrigues,C.H.M., Albanaz,A.T.S., Karmakar,M., Myung,Y., Xavier,J., Michanetzi,E.M., Portelli,S. and Ascher,D.B. (2019) Exploring protein supersecondary structure Through Changes in Protein Folding, Stability, and Flexibility. *Methods Mol. Biol.*, **1958**, 173–185.
15. Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline Mutations in the CDKN2B Tumor Suppressor Gene Predispose to Renal Cell Carcinoma. *Cancer Discov.*, **5**, 723–729.
16. Usher,J.L., Ascher,D.B., Pires,D.E., Milan,A.M., Blundell,T.L. and Ranganath,L.R. (2015) Analysis of HGD gene Mutations in Patients with Alkaptonuria from the United Kingdom: Identification of Novel Mutations. *JIMD Rep.*, **24**, 3–11.
17. Andrews,K.A., Vialard,L., Ascher,D.B., Pires,D.E.V., Bradshaw,N., Cole,T., Cook,J., Irving,R., Kumar,A., Lalloo,F. *et al.* (2016) Tumour risks and genotype-phenotype-proteotype analysis of patients with germline mutations in the succinate dehydrogenase subunit genes SDHB, SDHC, and SDHD. *Lancet*, **387**, 19–19.
18. Nemethova,M., Radvanszky,J., Kadasi,L., Ascher,D.B., Pires,D.E., Blundell,T.L., Porfirio,B., Mannoni,A., Santucci,A., Milucci,L. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur. J. Hum. Genet.*, **24**, 66–72.
19. Pires,D.E., Blundell,T.L. and Ascher,D.B. (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.*, **6**, 29575.
20. Albanaz,A.T.S., Rodrigues,C.H.M., Pires,D.E.V. and Ascher,D.B. (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin. Drug Discov.*, **12**, 553–563.
21. Casey,R.T., Ascher,D.B., Rattenberry,E., Izatt,L., Andrews,K.A., Simpson,H.L., Challis,B., Park,S.M., Bulusu,V.R., Lalloo,F. *et al.* (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol. Genet. Genomic Med.*, **5**, 237–250.
22. Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B., Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom. Med.*, **2**, 7.
23. Andrews,K.A., Ascher,D.B., Pires,D.E.V., Barnes,D.R., Vialard,L., Casey,R.T., Bradshaw,N., Adlard,J., Aylwin,S., Brennan,P. *et al.* (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J. Med. Genet.*, **55**, 384–394.
24. Hnizda,A., Fabry,M., Moriyama,T., Pachl,P., Kugler,M., Brinsa,V., Ascher,D.B., Carroll,W.L., Novak,P., Zaliova,M. *et al.* (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. *Leukemia*, **32**, 1393–1403.
25. Ascher,D.B., Spiga,O., Sekelska,M., Pires,D.E.V., Bernini,A., Tiezzi,M., Kralovicova,J., Borovska,I., Soltysova,A., Olsson,B. *et al.* (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.*, **27**, 888–902.
26. Bayley,J.P., Bausch,B., Rijken,J.A., van Hulsteijn,L.T., Jansen,J.C., Ascher,D., Pires,D.E.V., Hes,F.J., Hensen,E.F., Corssmit,E.P.M. *et al.* (2020) Variant type is associated with disease characteristics in SDHB, SDHC and SDHD-linked phaeochromocytoma-paraganglioma. *J. Med. Genet.*, **57**, 96–103.
27. Trezza,A., Bernini,A., Langella,A., Ascher,D.B., Pires,D.E.V., Sodi,A., Passerini,I., Pelo,E., Rizzo,S., Niccolai,N. *et al.* (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest Ophthalmol. Vis. Sci.*, **58**, 5320–5328.
28. Kano,F.S., Souza-Silva,F.A., Torres,L.M., Lima,B.A., Sousa,T.N., Alves,J.R., Rocha,R.S., Fontes,C.J., Sanchez,B.A., Adams,J.H. *et al.*

(2016) The presence, persistence and functional properties of Plasmodium vivax duffy binding protein II antibodies are influenced by HLA class II allelic variants. *PLoS Negl. Trop. Dis.*, **10**, e0005177.
29. Silvino,A.C., Costa,G.L., Araujo,F.C., Ascher,D.B., Pires,D.E., Fontes,C.J., Carvalho,L.H., Brito,C.F. and Sousa,T.N. (2016) Variation in human cytochrome *P*-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. *PLoS One*, **11**, e0160172.
30. White,R.R., Ponsford,A.H., Weekes,M.P., Rodrigues,R.B., Ascher,D.B., Mol,M., Selkirk,M.E., Gygi,S.P., Sanderson,C.M. and Artavanis-Tsakonas,K. (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. *PLoS Pathog*, **12**, e1005977.
31. Hawkey,J., Ascher,D.B., Judd,L.M., Wick,R.R., Kostoulias,X., Cleland,H., Spelman,D.W., Padiglione,A., Peleg,A.Y. and Holt,K.E. (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. *Microbial Genomics*, **4**, e000165.
32. Holt,K.E., McAdam,P., Thai,P.V.K., Thuong,N.T.T., Ha,D.T.M., Lan,N.N., Lan,N.H., Nhu,N.T.Q., Hai,H.T., Ha,V.T.N. *et al.* (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat Genet*, **50**, 849–856.
33. Karmakar,M., Globan,M., Fyfe,J.A.M., Stinear,T.P., Johnson,P.D.R., Holmes,N.E., Denholm,J.T. and Ascher,D.B. (2018) Analysis of a novel pncA mutation for susceptibility to Pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.*, **198**, 541–544.
34. Portelli,S., Phelan,J.E., Ascher,D.B., Clark,T.G. and Furnham,N. (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Sci. Rep.*, **8**, 15356.
35. Vedithi,S.C., Malhotra,S., Das,M., Daniel,S., Kishore,N., George,A., Arumugam,S., Rajan,L., Ebenezer,M., Ascher,D.B. *et al.* (2018) Structural implications of mutations conferring Rifampin resistance in Mycobacterium leprae. *Sci. Rep.*, **8**, 5016.
36. Karmakar,M., Rodrigues,C.H.M., Holt,K.E., Dunstan,S.J., Denholm,J. and Ascher,D.B. (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS One*, **14**, e0217169.
37. Chaitanya Vedithi,S., Rodrigues,C.H.M., Portelli,S., Skwark,M.J., Das,M., Ascher,D.B., Blundell,T.L. and Malhotra,S. (2020) Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of RNA polymerase in Mycobacterium leprae. *Comput. Struct. Biotechnol. J.*, **18**, 271–286.
38. Karmakar,M., Rodrigues,C.H.M., Horan,K., Denholm,J.T. and Ascher,D.B. (2020) Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci Rep*, **10**, 1875.
39. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
40. Popov,P., Bizin,I., Gromiha,M., A,K. and Frishman,D. (2019) Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure. *PLoS One*, **14**, e0219452.
41. Famiglietti,M.L., Estreicher,A., Gos,A., Bolleman,J., Gehant,S., Breuza,L., Bridge,A., Poux,S., Redaschi,N., Bougueleret,L. *et al.* (2014) Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.*, **35**, 927–935.
42. Jubb,H.C., Higueruelo,A.P., Ochoa-Montano,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.
43. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
44. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, **36**, D202–D205.
45. Wilkins,M.R., Gasteiger,E., Bairoch,A., Sanchez,J.C., Williams,K.L., Appel,R.D. and Hochstrasser,D.F. (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.*, **112**, 531–552.
46. Omasits,U., Ahrens,C.H., Muller,S. and Wollscheid,B. (2014) Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics*, **30**, 884–886.
47. Pieper,U., Webb,B.M., Dong,G.Q., Schneidman-Duhovny,D., Fan,H., Kim,S.J., Khuri,N., Spill,Y.G., Weinkam,P., Hammel,M. *et al.* (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **42**, D336–D346.

Appendix H

EasyVS: a user friendly web based tool for molecule library selection and structure-based virtual screening

Bioinformatics, 36(14), 2020, 4200–4202
doi: 10.1093/bioinformatics/btaa480
Advance Access Publication Date: 12 May 2020
Applications Note

Structural bioinformatics

EasyVS: a user-friendly web-based tool for molecule library selection and structure-based virtual screening

Douglas E. V. Pires [1,2,3,*,†], Wandré N. P. Veloso [4,†], YooChan Myung[2,3,†], Carlos H. M. Rodrigues[2,3,†], Michael Silk[2,3,†], Pâmela M. Rezende[5,†], Francislon Silva[5], Joicymara S. Xavier[5,6], João P. L. Velloso[5], Carlos H. da Silveira[4] and David B. Ascher[2,3,7,*,†]

[1]School of Computing and Information Systems, University of Melbourne, Melbourne 3010, Australia, [2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne 3004, Australia, [3]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne 3010, Australia, [4]Institute of Technological Sciences, Universidade Federal de Itajubá, Itabira 35903-087, Brazil, [5]Instituto René Rachou, Fundação Oswaldo Cruz, Belo Horizonte 30190-002, Brazil, [6]Instituto de Ciências Agrárias, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Unaí 38610-000, Brazil and [7]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK

*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.
Associate Editor: Yann Ponty

Received on December 31, 2018; revised on April 1, 2020; editorial decision on May 3, 2020; accepted on May 5, 2020

Abstract

Summary: EasyVS is a web-based platform built to simplify molecule library selection and virtual screening. With an intuitive interface, the tool allows users to go from selecting a protein target with a known structure and tailoring a purchasable molecule library to performing and visualizing docking in a few clicks. Our system also allows users to filter screening libraries based on molecule properties, cluster molecules by similarity and personalize docking parameters.

Availability and implementation: EasyVS is freely available as an easy-to-use web interface at http://biosig.unimelb.edu.au/easyvs.

Contact: douglas.pires@unimelb.edu.au or david.ascher@unimelb.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structure-based virtual screening has been widely and successfully used in early stages of drug development, aiding in the identification of potential hits and guiding further experimental validation (Cheng *et al.*, 2012). Molecular docking is one of the most widely used virtual screening approaches, which uses the three-dimensional structure of a target protein to predict the predominant binding mode of a small molecule with the target of interest. In this way, docking can be used to evaluate a large library of molecules and identify those most likely to interact with the target in the desired manner. This has been a powerful tool in the identification of initial hits, significantly reducing the chemical space to experimentally test and increasing the proportion of positive compounds being screened.

Significant improvements in docking protocols (Di Muzio *et al.*, 2017), scoring functions (Pires and Ascher, 2016) and molecule libraries (Sterling and Irwin, 2015), and the greater availability of computational power, have made virtual screening a more tractable and reliable hit identification strategy. Despite this, current virtual screening approaches typically require specialist computational and technical expertise.

In order to make virtual screening more friendly and accessible to a wider audience, we propose EasyVS (http://biosig.unimelb.edu.au/easyvs), a web-based, efficient and intuitive system that allows users to go from defining a protein structure and molecule library to performing docking in a few clicks. Our system allows users to optimize their screening library based on their properties (and define a chemical space of interest). Through a molecule clustering approach, we can more rapidly screen a larger chemical space, and present the top solutions to the user.

2 Platform description

Initially users are asked to define their target of interest by either uploading a structure of interest, or using the biological assembly of

a previous experimental structure deposited in the Protein Data Bank by providing the PDB accession code.

The selected structure is then analyzed using Ghecom (Kawabata, 2010) to identify druggable pockets. While by default the largest pocket is chosen for docking, users may select another pocket of interest for screening, and can refine the boundaries and docking parameters used (Supplementary Fig. S1). These parameters include box size and position (which can be set to any of the identified pockets and finely adjusted by the user) and depth of the search.

In the next step, users are asked to define the set of molecules to be assessed via docking. EasyVS currently supports the small molecule databases ChEMBL 24_1 (Gaulton *et al.*, 2017), HMDB 4.0 (Wishart *et al.*, 2018a), Drugbank 5.0 (Wishart *et al.*, 2018b), Maybridge (https://www.maybridge.com), Super Natural II (Banerjee *et al.*, 2015), Chembridge (Desai *et al.*, 2004) and Zinc15 (Sterling and Irwin, 2015), which together comprise over 16 million molecules.

There are many filters available for refining these molecular libraries, including by molecular weight, number of acceptors or donors of hydrogen, log*P* (or only selecting Lipinski's Ro5 molecules), fragments or natural products (Fig. 1A). Once the molecule library has been selected, users have the option to cluster molecules by similarity to improve screening performance. If users opt to perform clustering, one representative molecule from each group is randomly selected for docking. Users have the option to select the level of similarity used during clustering, which will change the number of clusters and, therefore, the number of molecules that will proceed for docking stages.

Molecule docking of the selected compound library is then performed using Autodock Vina (Trott and Olson, 2010). Users can also rapidly rescore selected poses using NNScore (Durrant and McCammon, 2011) or CSM-Lig (Pires and Ascher, 2016), analyze the intramolecular interactions using Arpeggio (Jubb et al., 2017) and predict pharmacokinetic properties of top hits using pkCSM (Pires *et al.*, 2015). While the docking is running, users can view the results in real time to analyze best poses of selected molecules (Fig. 1B) as well as include additional molecules for docking. Further exploration of the chemical space of top docked ligands is available by an additional round of virtual screening, using compounds that are structurally similar to the ligand of interest.

The EasyVS docking protocol was validated using two different benchmarks. We performed a redocking procedure for a selection of eight G-protein-coupled receptors (GPCR)-ligands complexes with available crystallographic structures. Ligands have been successfully redocked with an average root-mean-square deviation (RMSD) of 0.98 Å (Supplementary Table S1). We have also created decoy libraries using DUD-e (Mysinger *et al.*, 2012) for the same set of proteins considered for redocking. The docking scores for real ligands were considerably higher than those obtained for the decoys (*P*-value <0.001, Supplementary Table S2), demonstrating the robustness of the EasyVS docking protocol. We have evaluated the system's ability to process multiple submissions, demonstrating the system's responsiveness (Supplementary Fig. S2).

3 Conclusions

Here, we present EasyVS, a freely available, user-friendly platform for simplifying molecule library construction and docking. EasyVS allows users to choose molecules from well-established and diverse databases, including fragments, approved drugs and natural products, and perform the docking with just a few clicks. We also show EasyVS was successful in identifying GPCR ligands (Supplementary Materials) as a case study.

We believe this will be an invaluable tool for the exploratory stages of hit identification, allowing for the selection either stringent or very diverse sets of molecules for virtual screening and the intelligent assessment of different small molecule chemical spaces.

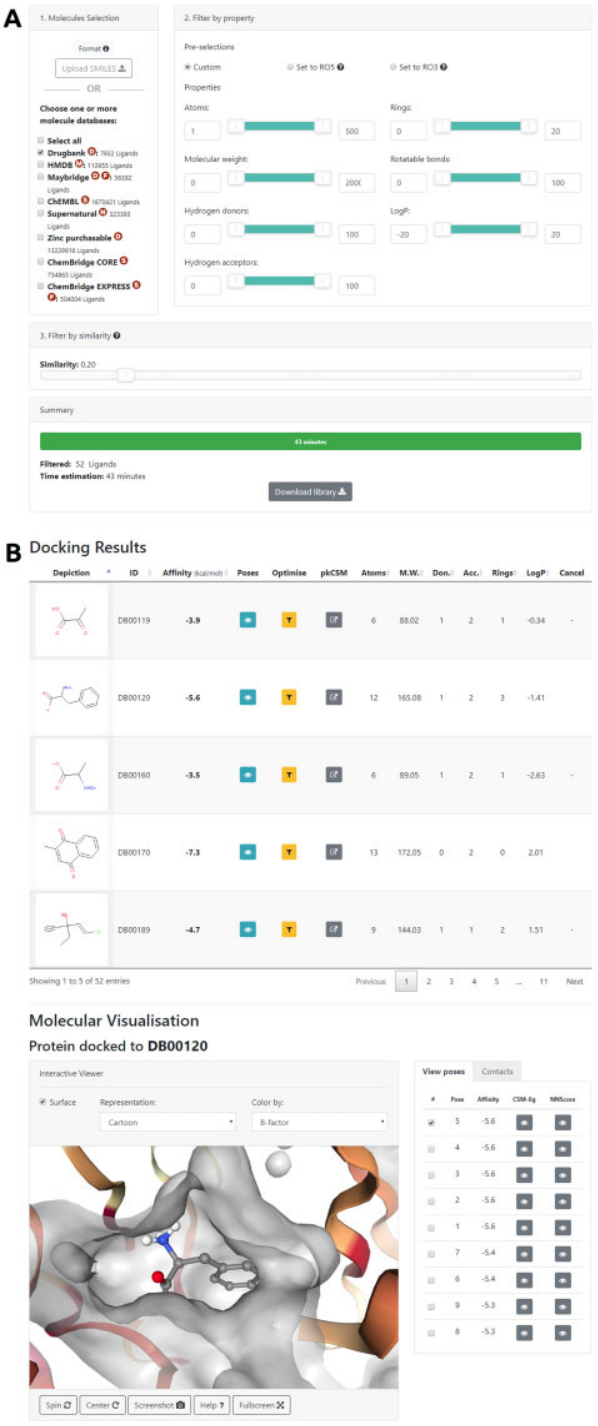


Fig. 1. EasyVS web interface. Once a protein target is selected, (**A**) users are prompted to select a molecule library from available databases and property filters. (**B**) Docking will be performed on the selected target/library set and best poses shown as an interactive molecule visualization

Funding

This work has been supported by the Melbourne Research Scholarships (to C.H.M.R and Y.M.); Medical Research Council (MRC) [MR/M026302/1 to D.B.A., D.E.V.P.]; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Finance Code 001 (to C.H.S.); National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (to D.E.V.P. and C.H.S.); Universidade Federal de Itajubá (to W.N.P.V. and C.H.S.). Supported in part by the Victorian Government's OIS Program.

Conflict of Interest: none declared.

References

Banerjee,P. *et al.* (2015) Super Natural II—a database of natural products. *Nucleic Acids Res.*, **43**, D935–D939.

Cheng,T. *et al.* (2012) Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.*, **14**, 133–141.

Desai,P.V. *et al.* (2004) Identification of novel parasitic cysteine protease inhibitors using virtual screening. The ChemBridge database. *J. Med. Chem.*, **47**, 6609–6615.

Di Muzio,E. *et al.* (2017) DockingApp: a user friendly interface for facilitated docking simulations with AutoDock Vina. *J. Comput. Aided Mol. Des.*, **31**, 213–218.

Durrant,J.D. and McCammon,J.A. (2011) NNScore 2.0: a neural-network receptor–ligand scoring function. *J. Chem. Inf. Model.*, **51**, 2897–2903.

Gaulton,A. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.

Jubb,H.C. *et al.* (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.

Kawabata,T. (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins Struct. Funct. Bioinf.*, **78**, 1195–1211.

Mysinger,M.M. *et al.* (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.*, **55**, 6582–6594.

Pires,D.E.V. and Ascher,D.B. (2016) CSM-lig: a web server for assessing and comparing protein–small molecule affinities. *Nucleic Acids Res.*, **44**, W557–W561.

Pires,D.E. *et al.* (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem.*, **58**, 4066–4072.

Sterling,T. and Irwin,J.J. (2015) ZINC 15–ligand discovery for everyone. *J. Inf. Model.*, **55**, 2324–2337.

Trott,O. and Olson,A.J. (2009) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.

Wishart,D.S. *et al.* (2018a) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.

Wishart,D.S. *et al.* (2018b) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.

Appendix I

Empirical ways to identify novel Bedaquiline resistance mutations in AtpE

PLOS ONE

RESEARCH ARTICLE

Empirical ways to identify novel Bedaquiline resistance mutations in AtpE

Malancha Karmakar[1,2,3,4], **Carlos H. M. Rodrigues**[2,4], **Kathryn E. Holt**[2], **Sarah J. Dunstan**[5], **Justin Denholm**[1,3], **David B. Ascher**[2,4,6]*

1 Victorian Tuberculosis Program, Melbourne Health, Victoria, Australia, **2** Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Victoria, Australia, **3** Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia, **4** Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia, **5** The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Victoria, Australia, **6** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

* david.ascher@unimelb.edu.au

Citation: Karmakar M, Rodrigues CHM, Holt KE, Dunstan SJ, Denholm J, Ascher DB (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS ONE 14(5): e0217169. https://doi.org/10.1371/journal.pone.0217169

Editor: Igor Mokrousov, St Petersburg Pasteur Institute, RUSSIAN FEDERATION

Received: January 31, 2019

Accepted: May 1, 2019

Published: May 29, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: M.K was funded by the Melbourne Research Scholarship. D.B.A was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), and a C. J. Martin Research Fellowship from the National

Abstract

Clinical resistance against Bedaquiline, the first new anti-tuberculosis compound with a novel mechanism of action in over 40 years, has already been detected in *Mycobacterium tuberculosis*. As a new drug, however, there is currently insufficient clinical data to facilitate reliable and timely identification of genomic determinants of resistance. Here we investigate the structural basis for *M. tuberculosis* associated bedaquiline resistance in the drug target, AtpE. Together with the 9 previously identified resistance-associated variants in AtpE, 54 non-resistance-associated mutations were identified through comparisons of bedaquiline susceptibility across 23 different mycobacterial species. Computational analysis of the structural and functional consequences of these variants revealed that resistance associated variants were mainly localized at the drug binding site, disrupting key interactions with bedaquiline leading to reduced binding affinity. This was used to train a supervised predictive algorithm, which accurately identified likely resistance mutations (93.3% accuracy). Application of this model to circulating variants present in the Asia-Pacific region suggests that current circulating variants are likely to be susceptible to bedaquiline. We have made this model freely available through a user-friendly web interface called SUSPECT-BDQ, StrUctural Susceptibility PrEdiCTion for bedaquiline (http://biosig.unimelb.edu.au/suspect_bdq/). This tool could be useful for the rapid characterization of novel clinical variants, to help guide the effective use of bedaquiline, and to minimize the spread of clinical resistance.

Introduction

Tuberculosis (TB) is the leading cause of infectious disease death worldwide, with over 10 million new cases and 1.6 million deaths in 2017 [1]. A disproportionate burden arises from the estimated 558,000 annual cases of rifampicin resistant TB (RR-TB) with 82% being multi-drug resistant (MDR), which is associated with lengthy, toxic therapy and high rates of mortality [1]. With limited therapeutic options available, especially for MDR-TB and extensively drug-

Health and Medical Research Council (NHMRC) of Australia (APP1072476). The Vietnam genomic dataset was funded by a NHMRC Australia grant (APP1056689) to SJD and KEH. This work was supported in part by the Victorian Government's OIS Program. No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

resistant (XDR) TB, the introduction of new treatment options is urgently required. Bedaquiline, a new anti-TB drug with a novel mechanism of action, targeting the c-ring of ATP synthase (AtpE) [2], was approved for treatment for MDR-TB in 2012 [3, 4]. This innovative drug is potent against both actively replicating and dormant bacilli and has been shown to increase culture conversion in patients with MDR-TB [5]. The use of bedaquiline has expanded considerably in recent years, and has been recommended for more routine use in MDR-TB regimens [6], however clinical failures have already been observed [7, 8]. This necessitates a better understanding of how variants result in resistance to aid in the early detection of resistance.

Phenotypic, and increasingly genotypic, drug susceptibility testing (DST) is recognized as essential for effective individualization of TB therapy. However, while progress has been made in strengthening laboratory diagnostics, the TB community is still struggling to build up laboratory networks with the needed capacity for routine culture and DST [7, 9]. The World Health Organization (WHO) has strongly urged the development of accurate and reproducible DST for bedaquiline and recommended that in the absence of specific DST, bedaquiline resistance should be monitored through MIC assessment [10] with resistance development evaluated in patients with treatment failure or relapse. Early characterization of drug resistance mutations would assist TB patient management and avoid treating individuals with ineffective toxic regimens [11, 12], but capacity for rapid genotypic prediction of bedaquiline resistance is limited by the identification of few known resistance associated variants [13].

In an era of rapidly expanding use of molecular technologies, including whole genome sequencing, tools for evaluating the impact of novel mutations are increasingly vital, particularly for drug resistance to novel and emerging medications such as bedaquiline. Though culture-based detection of resistance will remain the gold standard, *in silico* analyses can support informed decision-making. We have previously shown that the analysis of how variants can affect protein structure and function can be used to reliably characterize how variants lead to drug resistance [14–18]. Using this approach, we have shown that drug resistant mutations can be rapidly, accurately and pre-emptively predicted, guiding drug development [19–22] and clinical diagnosis [23].

In-vitro selection [24] and clinical studies [25] have shown that variants in the *atpE* gene can lead to bedaquiline resistance. To support rapid identification of potential bedaquiline resistance mutations, we considered whether structural information of the drug target could help guide clinical inference on genomic variants. Using a suite of well-established computational tools for characterizing the molecular consequences of mutations on protein structure and function, we have assessed the effects of mutations on the biophysical changes of AtpE folding, stability and on drug binding affinity. This was used to characterize how mutations in AtpE lead to resistance, and to train a predictive multilayer perceptron (feedforward artificial neural network) algorithm to characterize novel AtpE variants.

Methods

Data sets

Resistant variants from *in-vitro* selection studies were curated [13, 24, 26] along with a natural variant [4, 27] and used for model development. Susceptible variants were identified using a novel homology approach, where the genomes of all mycobacteria species sensitive to the drug [28] were aligned, therefore inferring that any present variants were likely to be susceptible. Clinically observed bedaquiline resistant *atpE* variants were curated from published reports [25]. The Vietnam dataset consists of whole genome sequences of 1635 *Mycobacterium tuberculosis* (*Mtb*) strains isolated from patients with pulmonary TB in Ho Chi Minh City, Vietnam.

The *Mtb* genome data is available in NCBI BioProject [ID: PRJNA355614; http://www.ncbi.nlm.nih.gov/bioproject/355614]. Details of the clinical study and the whole genome dataset are found in Thai et al [29] and Holt et al [15].

Homology modeling of AtpE

The structure of *Mtb* AtpE was modelled with MODELLER [30] using the experimental crystal structure of *Mycobacterium phlei* (*M. phlei*) AtpE (PDB ID: 4V1F). The model was then minimized in Prime and bedaquiline docked into the apo structure using Glide (Schrödinger Suite).

Modelling the biophysical consequences of missense variants

The structural consequences of the AtpE polymorphisms were assessed to account for all the potential effects of the mutations. The effects of mutations on protein folding and stability were assessed using SDM [31], mCSM-Stability [32] and DUET [33], and their effects on protein flexibility and conformation was predicted using normal mode analysis by DynaMut [34]. The effect of the difference on the protein-protein interactions between the protomers of AtpE were predicted using mCSM-PPI [32]. The effect of the changes on the binding affinity of bedaquiline towards AtpE were predicted using mCSM-Lig [35–37]. These approaches are novel machine-learning algorithms that use graph-based signatures to represent the structural and chemical environment of the wild-type 3D structure of a protein to quantitatively predict the effects of point mutations. Additionally, SNAP2 [38] was used to provide additional evolutionary based information.

Machine learning

To build the binary classifier, a multilayer perceptron neural network algorithm was trained, based on the implementation available through the Weka toolkit [39]. The resistant variants were up-sampled to create a more balanced model [40]. The training dataset constituted of 50 non-resistant associated variants and 5 resistant associated variants, while the blind test dataset constituted of 4 non-resistant associated variants and 4 resistant associated variants. To avoid over-biasing, the train and blind test dataset were non-redundant with respect to residue position. The model was trained and evaluated using jackknife [41] and leave-one-residue-position-out validation. The classification model was evaluated based on metrics, including the Area Under the ROC curve (AUC), precision and accuracy. Statistical analysis was performed using RStudio (version 3.1.1).

Webserver development

The server front-end was built using materialize CSS framework version 1.0.0, while the back-end was built in Python via the Flask framework (version 0.12.2). It is hosted on a Linux server running Apache.

Results

We used a structure-guided approach to understand the protein structure of the drug target AtpE and machine learning to build an empirical tool that could identify likely resistant mutations. The pipeline used to analyze the variants and train a multilayer perceptron neural network algorithm is shown in Fig 1.

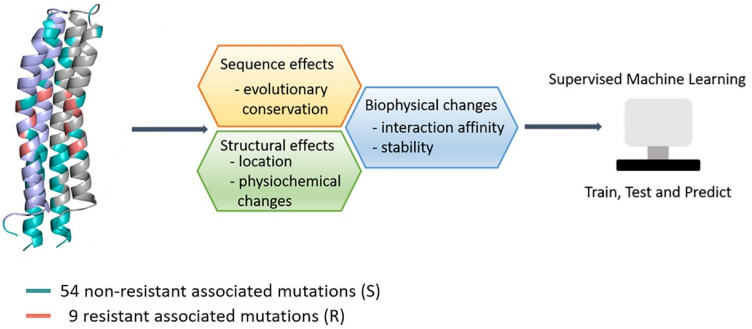


Fig 1. Methodology. This workflow highlights important steps in the methodology and how the main components of the algorithms are computed. In our analysis we used 54 non-resistant associated mutations and 9 resistant mutations for the biophysical analysis, followed by training and validation of our empirical model using a supervised machine learning algorithm.

https://doi.org/10.1371/journal.pone.0217169.g001

Structural information: The drug binding domain

A homology model of *Mtb* H37Rv AtpE was built using the existing experimental crystal structure of AtpE from *Mycobacterium phlei* (PDB ID: 4V1F) [42], which shares a high sequence identity with the *Mtb* protein (84.9%). The protomer model was an alpha helical hairpin structure comprising two membrane-spanning helices connected by a hydrophilic loop. The homo-oligomeric construct was built using the *M. phlei* structure as a guide, as the *Mtb* protein has been previously shown to assemble as a homo-nonamer [43] (Fig 2A and 2B). The cylindrical palisade model contained an internal hydrophobic cavity where phospholipid had been proposed to bind. The conserved proton binding residue (E61) was located sandwiched between adjacent protomers and equidistantly distributed along the center of the hydrophobic membrane bilayer.

The top docking poses of bedaquiline with the nonamer homology model identified a pose consistent with that observed in the *M. phlei* structure. The drug binding cleft was located at the interface of two protomers, with amino acid residues E61, A62, Y64, F65 from one protomer and I66 from the adjacent protomer defining the drug binding cleft. Analysis of the molecular interactions with Arpeggio [44] highlighted a strong network of polar interactions between the drug and AtpE (Fig 2C). Of particular interest, the diethylaminomethyl group of bedaquiline specifically interacted with the conserved proton binding residue E61, making

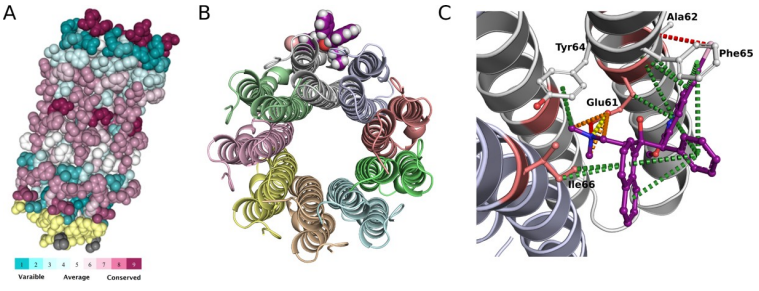


Fig 2. Structure and sequence information. (A) ConSurf analysis of AtpE (*M. tuberculosis*) where the evolutionary rates of conservation are color-coded on to the structure. (B) The experimental crystal structure of AtpE bound to Bedaquiline (purple). (C) The key molecular interaction between Bedaquiline (ball and stick representation; purple) and AtpE: ionic bond (yellow), π-interactions (green), proximal hydrogen bond (red) and weak polar van der Waal clashes (orange). The known resistance mutations are shown as salmon red (sticks) on the cartoon representation of the AtpE structure.

https://doi.org/10.1371/journal.pone.0217169.g002

tight ionic and hydrogen bonds with the carboxyl group of E61 (S1 Fig). In the docked model, bedaquiline also made strong π-interactions with residues Y64 and I66, and a hydrogen bond to A62.

Variant calling

We identified 9 previously published bedaquiline resistant non-synonymous single nucleotide variants (nsSNVs) from *in-vitro* selection experiments [4, 13, 24, 26]. To identify AtpE mutations not associated with drug resistance, we examined sequence variation amongst AtpE sequences from 23 mycobacterial species that have been shown to be phenotypically sensitive to the drug [27, 45–49] (Fig 3). Due to the high degree of sequence conservation across mycobacterial AtpE sequences (~ 66% sequence homology; Clustal Omega), variations between strains shown to be susceptible to bedaquiline were inferred to not be associated with drug resistance. Through comparison against the *Mtb* sequence (highlighted in yellow in Fig 3), 54 non-resistance-associated variants were identified (shown in teal in Fig 3).

Understanding the structural basis of resistance is important to facilitate the rapid identification of novel resistance variants, aiding efforts to minimize the rapid development of resistance [23]. The 54 non-resistance-associated variants ("S") and 9 resistant variants ("R") were mapped on the protein structure of AtpE (Fig 1). Most of the non-resistance-associated mutations were located on the N-terminal surface exposed inner loop of AtpE. Conserved regions (highlighted red in Fig 3) were evident, mainly on the C-terminal or the outer loop and embedded in the lipid bilayer of the membrane. All resistance-associated mutations were localized within 5 Å of the known drug binding site, which we refer to as the "resistance hotspot".

Structural and biophysical consequences of AtpE variants

The resistant associated variants were all predicted by SNAP2 [38] to be more functionally deleterious than the non-resistance associated variants, reflecting the resistant associated variants are in a more conserved region of the protein. In order to better understand the molecular consequences of the mutations on AtpE structure and function, the mutations were analyzed in the context of both the apo and complexed protomeric structures. The impact of resistant and non-resistant associated mutations on protein folding, stability and conformation were assessed using SDM [31], mCSM-Stability [32], DUET [33] and DynaMut [34]. The effect of the variants on the affinity of the protomers to form the cylindrical palisade homo-oligomer were examined using mCSM-PPI [32], and the effect of the variants on the binding affinity for bedaquiline were assessed using mCSM-Lig [37].

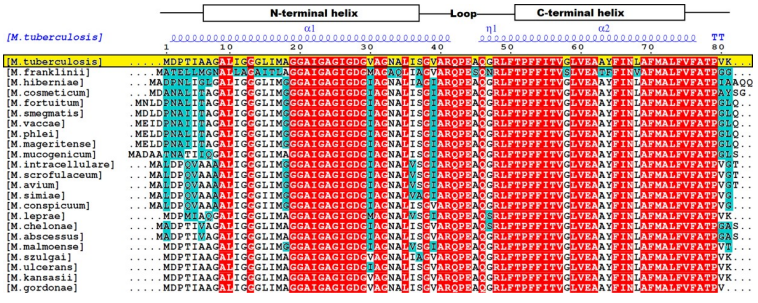


Fig 3. Non-resistant associated variant assignment. This image highlights the sequence alignment of 23 mycobacterial species sensitive to Bedaquiline. Residues that were different to the reference *M.tuberculosis* sequence (in yellow) are highlighted in teal, and were chosen as non-resistant associated variants for building the empirical model. The conserved residues are shown in red. The secondary structure of the AtpE protein is shown above the sequences in blue (α = alpha helix, η = loop). This image was created using ESPript 3 [56].

https://doi.org/10.1371/journal.pone.0217169.g003

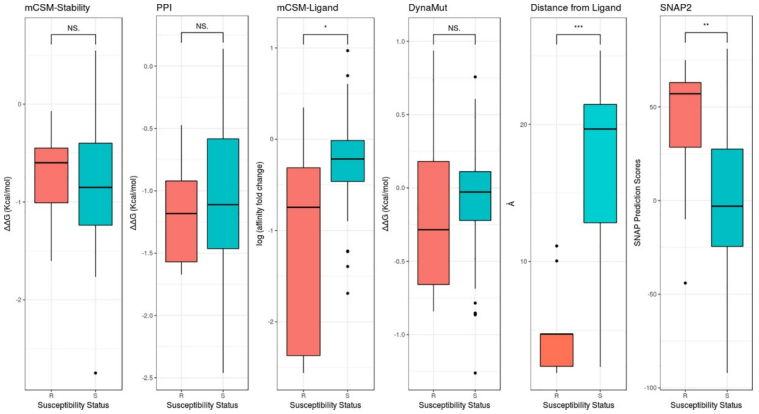


Fig 4. PCA analysis. Boxplot representation of all the features used to build the predictive model. The resistant associated mutations (R) are represented as red and the non-resistant associated mutations (S) as teal. (* $p<0.05$, ** $p<0.005$, *** $p<0.0001$, NS $p>0.5$ by Welch two sample t-test).

https://doi.org/10.1371/journal.pone.0217169.g004

Analysis of the variant effects on protomer stability and the formation of the cylindrical palisade did not reveal statistically significant differences between resistant and non-resistance-associated variants (Fig 4). This is consistent with recent work that showed in order to minimize fitness costs, resistant associated variants in drug targets tended to have mild effects on protein stability [50]. The largest destabilizing effect observed amongst the resistance-associated variants using mCSM-Stability and DUET was for the conservative mutation E61D ($\Delta\Delta G$ = -1.1 Kcal/mol), however normal mode analysis by DynaMut suggested that the E61D mutation would not destabilize the structure and was only associated with mild conformational changes (S1 Fig). Examination of residue conservation across 150 homologous sequences using ConSurf [51] showed the equivalent residue position in many species was an Asp, suggesting its introduction is unlikely to have a large structural or functional effect.

While all nine resistant variants were within 5 Å of the ligands, five in particular, A63M, A63P, E61D, L59V and I66M, were within 2.5 Å and making direct interactions with bedaquiline. Modelling of these mutations revealed that most of them would result in complete loss of these intermolecular interactions (S2 Fig). For example, E61 upon mutation to Asp would result in loss of these strong ionic and hydrogen bonds with bedaquiline. Interestingly, the mutation of I66 to Met and L59 to Val mutation revealed the formation of new interactions, although the overall binding affinity was predicted to be lower by CSM-lig. Most of the non-resistant associated variants were located distal to the bedaquiline binding site.

Analysis of predicted changes in bedaquiline binding affinity upon mutation using mCSM-Lig revealed a significant difference between variants associated with resistance or not associated with resistance (Fig 4). The non-resistance associated variants were associated with mild mCSM-Lig predicted changes in bedaquiline binding affinity (average of -0.25 log affinity fold change). This would be consistent with the mutations leading to minimal change in, or even increasing, drug binding affinity. The average predicted log fold change in binding affinity obtained for the 9 resistant mutations, by contrast, was -1.29 log affinity fold change, indicating that they would likely disrupt bedaquiline binding. Among them, all four D28 resistant variants were predicted to the largest destabilising effect on bedaquiline binding (-2.5 log affinity fold change on average). D28 is positioned on the inner helix of the protomer and is 4.7 Å from the drug binding site. When D28 was substituted with either Ala or Gly, a loss in inter-helical interactions and a gain in flexibility was observed, and when substituted to Pro and Val it led to a gain in intra-molecular interactions and rigidification of the AtpE structures (S2 Fig).

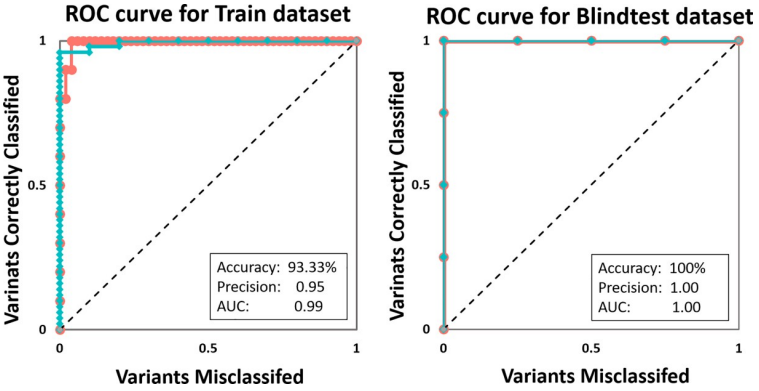


Fig 5. Evaluation metric. The ROC curve shows that using the structural and functional consequences of the variants, we were able to accurately identify resistant (red) and non-resistant associated (teal) variants.

https://doi.org/10.1371/journal.pone.0217169.g005

Machine learning algorithm: Multilayer perceptron network

Building on this structural analysis, we tested whether these structural features could be used to train a supervised machine learning algorithm capable of accurately predicting resistant associated variants. To avoid over-training, the 54 non-resistant and 9 resistant variants were split into a training and blind test dataset. Our training dataset constituted of 50 non-resistant associated variants and 5 resistant associated variants (A63V, A63P, I66M, L59V, E61D). Due to the small sample size, to balance the dataset, the resistant variants in the training dataset were oversampled (duplicated). The remaining 4 resistant (all D28 mutations) and 4 non-resistant associated (I11L, L15T, A34Q and A45S) variants in the blind test were positioned non-redundant with those in the training.

A list of features tested in method development is described in S1 Table. As discussed above, the features that best distinguished between the classes include distance from ligand binding site ("Distance from Ligand", $p < 0.0001$), mCSM-Lig ($p = 0.026$) and SNAP2 ($p < 0.0001$) (Fig 4). Using jackknife and leave-one-residue-position-out validation, models trained using multilayer perceptron neural networks yielded the strongest balanced performance. The final model correctly classified 93.33% and 100% of variants in the training and blind test datasets respectively (Fig 5, Table 1). The comparative performance across iterative non-redundant blind datasets suggested that the model was not over-fitted.

The classifier revealed that variants with mild effects on protein stability and conformation (DynaMut < 0.28 Kcal/mol and DUET < -1.65 Kcal/mol), located close to the docked bedaquiline (distance from ligand < 6.36 Å) were likely to be associated with resistance. A closer examination of the four incorrectly classified non-resistant associated variant in the train dataset revealed that three of them, G58S, A63T and L68V, were positioned very close to the bedaquiline binding site (< 2.5 Å) and N33A had a large predicted change in binding affinity (-1.4 log affinity fold change); indicating that these mutations might have direct consequences on bedaquiline binding.

Table 1. Evaluation metrics of the train and blind test dataset.

Multilayer Perceptron (MLP)	Precision score	Recall	F-measure	ROC area	PRC area
Train Dataset	0.952	0.933	0.938	0.970	0.967
Blind test Dataset	1.000	1.000	1.000	1.000	1.000

https://doi.org/10.1371/journal.pone.0217169.t001

Clinically identified resistance associated variants

Using a model trained without the D28 variants, we analyzed the recently reported clinical *atpE* bedaquiline resistant variants [25]. Both D28N and A63V were both predicted by the model to lead to bedaquiline resistance, consistent with the clinical data. Looking at these mutations within the structure, the mutation at D28 would disrupt interactions made by the wild-type residue to bedaquiline, consistent with the mCSM-Lig predictions that it would lead to a significant reduction in ligand binding affinity (S3 Fig; -1.87 log affinity fold change). Interestingly, while A63 did not make interactions directly with bedaquiline, the mutation to Val would lead to steric clashes with the bound ligand and prevent bedaquiline binding (S3 Fig).

Vietnam data analysis

We also used this approach to predict the sensitivity of two *atpE* nsSNVs, I16V and P52L, identified through whole genomic sequencing of *Mtb* strains isolated from 1635 TB patients in Vietnam [15]. The predictive tool classified the reported nsSNPs to be non-resistant associated variants. These variants were located approximately 10 Å away from the bedaquiline binding site, and mutations at these residues were not predicted to disrupt any interactions with bedaquiline (S4 Fig). As these samples had been collected from patients that had not been administered bedaquiline, it provided confidence that in our large analysis of patients in Vietnam there were no circulating strains likely to be resistant to bedaquiline.

SUSPECT-BDQ webserver

We have implemented SUSPECT-BDQ as a user-friendly, freely available web server http://biosig.unimelb.edu.au/suspect_bdq/. SUSPECT-BDQ provides two different input options. The "Single Mutation" option allows users to predict whether a mutation will be characterized as either Resistant or Susceptible. For this option, the server requires the point mutation to be specified as a text string containing the wild-type residue one-letter code, its corresponding position on the structure and the mutant one-letter code. The "Mutation List" option allows the user to upload a file with a list of mutations in a file for batch processing. In order to assist users to submit their mutations for analysis, sample submission entries are available for both input options and a help page is also available via the top navigation bar.

For the "Single Mutation" option, the web server displays the prediction outcome of SUSPECT-BDQ alongside with details of the user input data, information on the residue environment and parameters used on the prediction (S5 Fig). In addition, an interactive 3D viewer, built using NGL [52] allows for analysis of non-covalent inter-residue interactions for the position specified in the input calculated with Arpeggio [44] for wild-type and mutant structures. For the "Mutation List" option, the results are summarized in a downloadable table from which users can access details for each single mutation. A 3D viewer is also shown and each wild-type residue from the input list is colored according to the predicted effect.

Discussion

Early genomic detection of resistance is crucial for tailoring individual therapy and preventing the onward transmission of resistant infection. This is especially of importance to limit the spread of resistance to bedaquiline, one of the few treatment options for XDR-TB. While significant progress has been made in terms of innovative tools to understand and quantify the different range of effects in which a mutation or a set of mutations can give rise to a drug-resistant phenotype, a gap still exists when integrating these predictions and drawing conclusions

regarding causality and the strength of associations observed. This is compounded by the need for detailed information regarding the system/protein. The availability of scalable, effective computational methods to assess mutational effects creates new opportunities for developing integrated approaches and deciphering complex genomic background patterns, shedding light on their role in the emergence of a given phenotype and molecular mechanisms of action [19].

Here we have used a computational approach to better understand the molecular mechanism of drug resistance within the context of the protein's 3D structure. A machine learning algorithm was used to build a predictive tool which could pre-emptively determine novel bedaquiline resistant mutations within *atpE*. We began our investigation by studying the interaction dynamics between the c-ring of ATP synthase bound to bedaquiline. The correlations of conformational changes and Gibb's free energy provided novel molecular insights into how resistance variants affected bedaquiline binding but led to minimal disruption of protein folding and dynamics. Mapping of all the mutations on the crystal structure helped us identify the "mutational hotspot" for AtpE, which was in proximity to the drug binding site. We saw that resistance associated variants were more likely to be located within this resistance hotspot, and lead to a significant disruption in bedaquiline binding. Interestingly, the characterized resistant variants did not lead to large changes in protein folding, stability or oligomeric state, which would impose a larger fitness penalty [50].

This *in silico* biophysical information was used to build a predictive algorithm that accurately identified resistant mutations. We then prepared a comprehensive mutational dataset that contained the predictions of all possible mutations in AtpE, which we have made available through a web-based interface: SUSPECT_BDQ (http://biosig.unimelb.edu.au/suspect_bdq/). These analyses highlight the power of considering the structural environment of a mutation to understand the molecular and biological consequences [53]. As a relatively novel drug, there is still a paucity of reliable information regarding resistance mutations. While limited by the relatively small available datasets, repeated stratified non-redundant blind testing revealed the model was very robust. This associative approach thus helped us establish a set of guidelines which adds to the missing information in the database for new TB drugs like bedaquiline. It also provides a molecular understanding of how variants in AtpE affect ligand binding, leading to resistance, providing insight to guide development of second-generation inhibitors.

We intend further development of this tool through expanded genomic targets, and evaluation using additional clinical isolates. In particular we intend to extend SUSPECT_BDQ to include non-target based resistance to bedaquiline, which has been linked to mutations in *Rv0678* [54], a transcriptional repressor of the gene encoding the MmpS5-MmpL5 efflux pump, and *pepQ* (*Rv2535c*) [55], a putative Xaa-Pro aminopeptidase. Both are associated with low-level of resistance and therefore we did not include them in the study. However, low level resistance may have clinical significance in some settings, and future work will further evaluate other potentially important loci. Additionally, testing this tool on further clinical isolates will enhance the efficiency of the tool to predict the consequences of novel mutations.

Conclusion

This novel computational approach can enhance the impact of genome sequencing in identifying and characterizing variants more accurately and may therefore assist in guiding optimal usage of bedaquiline. The results obtained from our empirical tool is promising and should help facilitate routine genotypic drug susceptibility testing for bedaquiline and stimulate further research to help avoid the emergence of resistance to this new treatment through early detection.

Supporting information

S1 Table. The list of different features used to build the empirical model for predicting novel resistance associated mutations in bedaquiline.
(PDF)

S1 Fig. Detailed molecular interactions between the key proton binding residue E61, and upon its mutation to Asp, with bedaquiline. The wild-type residue is shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation). Hydrogen bonds are shown as orange dashes and ionic bond in yellow.
(TIF)

S2 Fig. Images of intermolecular interactions made by the wild-type residue (shown as cyan) and the mutant amino acid (shown as salmon red). Hydrogen bonds are shown in red, halogen bonds in blue, ionic bonds in yellow, hydrophobic bonds in green, π bonds in grey.
(TIF)

S3 Fig. Detailed molecular interactions between two clinically observed bedaquiline resistant variants, with the drug. The wild type residue is shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation). Halogen bonds are represented in blue dashes (amide-amide interaction) and π-bond as grey dashes.
(TIF)

S4 Fig. The localization of two circulating *atpE* variants relative to the bedaquiline binding pocket. The wild type residues are shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation).
(TIF)

S5 Fig. SUSPECT-BDQ webserver. Web-server results page for a single point mutation prediction. The predicted outcome is shown alongside with complementary information on the submitted mutation. An interactive 3D viewer allows for analysis of non-covalent interactions for both the wild type and mutant residue. In both cases controllers are provided in order to hide or show specific interactions and customize molecule representation.
(TIF)

Author Contributions

Conceptualization: Justin Denholm, David B. Ascher.

Data curation: Malancha Karmakar, Kathryn E. Holt, Sarah J. Dunstan, David B. Ascher.

Formal analysis: Malancha Karmakar, Kathryn E. Holt, Sarah J. Dunstan, Justin Denholm, David B. Ascher.

Funding acquisition: David B. Ascher.

Investigation: Malancha Karmakar, David B. Ascher.

Methodology: Malancha Karmakar, David B. Ascher.

Project administration: David B. Ascher.

Resources: Kathryn E. Holt.

Software: Carlos H. M. Rodrigues.

Supervision: David B. Ascher.

Validation: Malancha Karmakar, Kathryn E. Holt, Sarah J. Dunstan, Justin Denholm, David B. Ascher.

Writing – original draft: Malancha Karmakar, Carlos H. M. Rodrigues.

Writing – review & editing: Kathryn E. Holt, Sarah J. Dunstan, Justin Denholm, David B. Ascher.

References

1. WHO. Global Tuberculosis Report: Executive Summary. 2018; WHO/CDS/TB/2018.25.
2. Hards K, Robson JR, Berney M, Shaw L, Bald D, Koul A, et al. Bactericidal mode of action of bedaquiline. Journal of Antimicrobial Chemotherapy. 2015; 70(7):2028–37. https://doi.org/10.1093/jac/dkv054 PMID: 25754998
3. Koul A, Dendouga N, Vergauwen K, Molenberghs B, Vranckx L, Willebrords R, et al. Diarylquinolines target subunit c of mycobacterial ATP synthase. Nat Chem Biol. 2007; 3(6):323–4. Epub 2007/05/15. https://doi.org/10.1038/nchembio884 PMID: 17496888.
4. Petrella S, Cambau E, Chauffour A, Andries K, Jarlier V, Sougakoff W. Genetic basis for natural and acquired resistance to the diarylquinoline R207910 in mycobacteria. Antimicrob Agents Chemother. 2006; 50(8):2853–6. Epub 2006/07/28. https://doi.org/10.1128/AAC.00244-06 PMID: 16870785; PubMed Central PMCID: PMCPMC1538646.
5. Field SK. Bedaquiline for the treatment of multidrug-resistant tuberculosis: great promise or disappointment? Therapeutic Advances in Chronic Disease. 2015; 6(4):170–84. https://doi.org/10.1177/2040622315582325 PMC4480545. PMID: 26137207
6. WHO. Rapid Communication: Key changes to treatment of multidrug- and rifampicin-resistant tuberculosis (MDR/RR-TB). 2018. http://www.who.int/tb/publications/2018/rapid_communications_MDR/en/.
7. Salfinger M, Migliori GB. Bedaquiline: 10 years later, the drug susceptibility testing protocol is still pending. The European respiratory journal. 2015; 45(2):317–21. Epub 2015/02/06. https://doi.org/10.1183/09031936.00199814 PMID: 25653264.
8. Hoffmann H, Kohl TA, Hofmann-Thiel S, Merker M, Beckert P, Jaton K, et al. Delamanid and Bedaquiline Resistance in Mycobacterium tuberculosis Ancestral Beijing Genotype Causing Extensively Drug-Resistant Tuberculosis in a Tibetan Refugee. American journal of respiratory and critical care medicine. 2016; 193(3):337–40. Epub 2016/02/02. https://doi.org/10.1164/rccm.201502-0372LE PMID: 26829425.
9. Hoffmann H, Hofmann-Thiel S, Merker M, Kohl TA, Niemann S. Reply: Call for Regular Susceptibility Testing of Bedaquiline and Delamanid. American journal of respiratory and critical care medicine. 2016; 194(9):1171–2. Epub 2016/11/01. https://doi.org/10.1164/rccm.201605-1065LE PMID: 27797620.
10. WHO. The Use of Bedaquiline in the Treatment of Multidrug-Resistant Tuberculosis. 2013.
11. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome medicine. 2015; 7(1):51. Epub 2015/05/29. https://doi.org/10.1186/s13073-015-0164-0 PMID: 26019726; PubMed Central PMCID: PMCPMC4446134.
12. Nguyen TVA, Anthony RM, Banuls AL, Nguyen TVA, Vu DH, Alffenaar JC. Bedaquiline Resistance: Its Emergence, Mechanism, and Prevention. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2018; 66(10):1625–30. Epub 2017/11/11. https://doi.org/10.1093/cid/cix992 PMID: 29126225.
13. Segala E, Sougakoff W, Nevejans-Chauffour A, Jarlier V, Petrella S. New mutations in the mycobacterial ATP synthase: new insights into the binding of the diarylquinoline TMC207 to the ATP synthase C-ring structure. Antimicrob Agents Chemother. 2012; 56(5):2326–34. Epub 2012/02/23. https://doi.org/10.1128/AAC.06154-11 PMID: 22354303; PubMed Central PMCID: PMCPMC3346594.
14. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, et al. Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom. 2018. Epub 2018/03/17. https://doi.org/10.1099/mgen.0.000165 PMID: 29547094; PubMed Central PMCID: PMCPMC5885017.
15. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in

Vietnam. Nature genetics. 2018; 50(6):849–56. Epub 2018/05/23. https://doi.org/10.1038/s41588-018-0117-9 PMID: 29785015.

16. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC medicine. 2016; 14:31. Epub 2016/03/24. https://doi.org/10.1186/s12916-016-0575-9 PMID: 27005572; PubMed Central PMCID: PMCPMC4804620.

17. Pires DE, Chen J, Blundell TL, Ascher DB. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Scientific reports. 2016; 6:19848. Epub 2016/01/23. https://doi.org/10.1038/srep19848 PMID: 26797105; PubMed Central PMCID: PMCPMC4726175.

18. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, et al. Structural Implications of Mutations Conferring Rifampin Resistance in Mycobacterium leprae. Scientific reports. 2018; 8(1):5016. Epub 2018/03/24. https://doi.org/10.1038/s41598-018-23423-1 PMID: 29567948; PubMed Central PMCID: PMCPMC5864748.

19. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert opinion on drug discovery. 2017; 12(6):553–63. Epub 2017/05/12. https://doi.org/10.1080/17460441.2017.1322579 PMID: 28490289.

20. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, et al. Essential but Not Vulnerable: Indazole Sulfonamides Targeting Inosine Monophosphate Dehydrogenase as Potential Leads against Mycobacterium tuberculosis. ACS infectious diseases. 2017; 3(1):18–33. Epub 2016/10/06. https://doi.org/10.1021/acsinfecdis.6b00103 PMID: 27704782; PubMed Central PMCID: PMCPMC5972394.

21. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, et al. The Inosine Monophosphate Dehydrogenase, GuaB2, Is a Vulnerable New Bactericidal Drug Target for Tuberculosis. ACS infectious diseases. 2017; 3(1):5–17. Epub 2016/10/12. https://doi.org/10.1021/acsinfecdis.6b00102 PMID: 27726334; PubMed Central PMCID: PMCPMC5241705.

22. Trapero A, Pacitto A, Singh V, Sabbah M, Coyne AG, Mizrahi V, et al. Fragment-Based Approach to Targeting Inosine-5'-monophosphate Dehydrogenase (IMPDH) from Mycobacterium tuberculosis. Journal of medicinal chemistry. 2018; 61(7):2806–22. Epub 2018/03/17. https://doi.org/10.1021/acs.jmedchem.7b01622 PMID: 29547284; PubMed Central PMCID: PMCPMC5900554.

23. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, et al. Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy. American journal of respiratory and critical care medicine. 2018; 198(4):541–4. Epub 2018/04/26. https://doi.org/10.1164/rccm.201712-2572LE PMID: 29694240; PubMed Central PMCID: PMCPMC6118032.

24. Huitric E, Verhasselt P, Koul A, Andries K, Hoffner S, Andersson DI. Rates and mechanisms of resistance development in Mycobacterium tuberculosis to a novel diarylquinoline ATP synthase inhibitor. Antimicrob Agents Chemother. 2010; 54(3):1022–8. Epub 2009/12/30. https://doi.org/10.1128/AAC.01611-09 PMID: 20038615; PubMed Central PMCID: PMCPMC2825986.

25. Zimenkov DV, Nosova EY, Kulagina EV, Antonova OV, Arslanbaeva LR, Isakova AI, et al. Examination of bedaquiline- and linezolid-resistant Mycobacterium tuberculosis isolates from the Moscow region. The Journal of antimicrobial chemotherapy. 2017; 72(7):1901–6. Epub 2017/04/08. https://doi.org/10.1093/jac/dkx094 PMID: 28387862.

26. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science (New York, NY). 2005; 307(5707):223–7. Epub 2004/12/14. https://doi.org/10.1126/science.1106753 PMID: 15591164.

27. Huitric E, Verhasselt P, Andries K, Hoffner SE. In vitro antimycobacterial spectrum of a diarylquinoline ATP synthase inhibitor. Antimicrob Agents Chemother. 2007; 51(11):4202–4. Epub 2007/08/22. https://doi.org/10.1128/AAC.00181-07 PMID: 17709466; PubMed Central PMCID: PMCPMC2151410.

28. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011; 7:539. Epub 2011/10/13. https://doi.org/10.1038/msb.2011.75 PMID: 21988835; PubMed Central PMCID: PMCPMC3261699.

29. Thai PVK, Ha DTM, Hanh NT, Day J, Dunstan S, Nhu NTQ, et al. Bacterial risk factors for treatment failure and relapse among patients with isoniazid resistant tuberculosis. BMC Infectious Diseases. 2018; 18(1):112. https://doi.org/10.1186/s12879-018-3033-9 PMID: 29510687

30. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology. 1993; 234(3):779–815. Epub 1993/12/05. https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673.

31. Worth CL, Preissner R, Blundell TL. SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic acids research. 2011; 39(Web Server issue):W215–22. Epub 2011/05/20. https://doi.org/10.1093/nar/gkr363 PMID: 21593128; PubMed Central PMCID: PMCPMC3125769.

32. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics (Oxford, England). 2014; 30(3):335–42. Epub 2013/11/28. https://doi.org/10.1093/bioinformatics/btt691 PMID: 24281696; PubMed Central PMCID: PMCPMC3904523.

33. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic acids research. 2014; 42(Web Server issue): W314–9. Epub 2014/05/16. https://doi.org/10.1093/nar/gku411 PMID: 24829462; PubMed Central PMCID: PMCPMC4086143.

34. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic acids research. 2018; 46(W1):W350–w5. Epub 2018/05/03. https://doi.org/10.1093/nar/gky300 PMID: 29718330; PubMed Central PMCID: PMCPMC6031064.

35. Pires DE, Ascher DB. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic acids research. 2016; 44(W1):W557–61. Epub 2016/05/07. https://doi.org/10.1093/nar/gkw390 PMID: 27151202; PubMed Central PMCID: PMCPMC4987933.

36. Pires DE, Blundell TL, Ascher DB. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic acids research. 2015; 43(Database issue):D387–91. Epub 2014/10/18. https://doi.org/10.1093/nar/gku966 PMID: 25324307; PubMed Central PMCID: PMCPMC4384026.

37. Pires DE, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Scientific reports. 2016; 6:29575. Epub 2016/07/08. https://doi.org/10.1038/srep29575 PMID: 27384129; PubMed Central PMCID: PMCPMC4935856.

38. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC genomics. 2015; 16 Suppl 8:S1. Epub 2015/06/26. https://doi.org/10.1186/1471-2164-16-s8-s1 PMID: 26110438; PubMed Central PMCID: PMCPMC4480835.

39. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009; 11(1):10–8. https://doi.org/10.1145/1656274.1656278

40. Provost F. Machine learning from imbalanced data sets 101. Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets. 2000. citeulike-article-id:7616988.

41. Wager S, Hastie T, Efron B. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. Journal of machine learning research: JMLR. 2014; 15(1):1625–51. Epub 2015/01/13. PMID: 25580094; PubMed Central PMCID: PMCPMC4286302.

42. Preiss L, Langer JD, Yildiz O, Eckhardt-Strelau L, Guillemont JE, Koul A, et al. Structure of the mycobacterial ATP synthase Fo rotor ring in complex with the anti-TB drug bedaquiline. Science advances. 2015; 1(4):e1500106. Epub 2015/11/26. https://doi.org/10.1126/sciadv.1500106 PMID: 26601184; PubMed Central PMCID: PMCPMC4640650.

43. Lu P, Lill H, Bald D. ATP synthase in mycobacteria: special features and implications for a function as drug target. Biochim Biophys Acta. 2014; 1837(7):1208–18. Epub 2014/02/12. https://doi.org/10.1016/j.bbabio.2014.01.022 PMID: 24513197.

44. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. Journal of molecular biology. 2017; 429(3):365–71. Epub 2016/12/15. https://doi.org/10.1016/j.jmb.2016.12.004 PMID: 27964945; PubMed Central PMCID: PMCPMC5282402.

45. Aguilar-Ayala DA, Cnockaert M, Andre E, Andries K, Gonzalez YMJA, Vandamme P, et al. In vitro activity of bedaquiline against rapidly growing nontuberculous mycobacteria. Journal of medical microbiology. 2017; 66(8):1140–3. Epub 2017/07/28. https://doi.org/10.1099/jmm.0.000537 PMID: 28749330; PubMed Central PMCID: PMCPMC5817190.

46. Chahine EB, Karaoui LR, Mansour H. Bedaquiline: a novel diarylquinoline for multidrug-resistant tuberculosis. The Annals of pharmacotherapy. 2014; 48(1):107–15. Epub 2013/11/22. https://doi.org/10.1177/1060028013504087 PMID: 24259600.

47. Ji B, Chauffour A, Andries K, Jarlier V. Bactericidal activities of R207910 and other newer antimicrobial agents against Mycobacterium leprae in mice. Antimicrob Agents Chemother. 2006; 50(4):1558–60. Epub 2006/03/30. https://doi.org/10.1128/AAC.50.4.1558-1560.2006 PMID: 16569884; PubMed Central PMCID: PMCPMC1426933.

48. Ji B, Lefrancois S, Robert J, Chauffour A, Truffot C, Jarlier V. In vitro and in vivo activities of rifampin, streptomycin, amikacin, moxifloxacin, R207910, linezolid, and PA-824 against Mycobacterium ulcerans. Antimicrob Agents Chemother. 2006; 50(6):1921–6. Epub 2006/05/26. https://doi.org/10.1128/AAC.00052-06 PMID: 16723546; PubMed Central PMCID: PMCPMC1479135.

49. Pang Y, Zheng H, Tan Y, Song Y, Zhao Y. In Vitro Activity of Bedaquiline against Nontuberculous Mycobacteria in China. Antimicrob Agents Chemother. 2017; 61(5). Epub 2017/03/01. https://doi.org/10.1128/aac.02627-16 PMID: 28242674; PubMed Central PMCID: PMCPMC5404590.

50. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Scientific reports. 2018; 8(1):15356–. https://doi.org/10.1038/s41598-018-33370-6 PMID: 30337649.

51. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic acids research. 2016; 44(W1):W344–50. Epub 2016/05/12. https://doi.org/10.1093/nar/gkw408 PMID: 27166375; PubMed Central PMCID: PMCPMC4987940.

52. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics (Oxford, England). 2018; 34(21):3755–8. Epub 2018/06/01. https://doi.org/10.1093/bioinformatics/bty419 PMID: 29850778; PubMed Central PMCID: PMCPMC6198858.

53. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochemical Society transactions. 2017; 45(2):303–11. Epub 2017/04/15. https://doi.org/10.1042/BST20160422 PMID: 28408471; PubMed Central PMCID: PMCPMC5390495.

54. Bloemberg GV, Gagneux S, Böttger EC, Keller PM, Stuckia D, Trauner A, et al. Acquired Resistance to Bedaquiline and Delamanid in Therapy for Tuberculosis. New England Journal of Medicine. 2015; 373(20):1986–8. https://doi.org/10.1056/NEJMc1505196 PMID: 26559594. Language: English. Entry Date: 20151121. Revision Date: 20161125. Publication Type: case study. Journal Subset: Biomedical.

55. Almeida D, Ioerger T, Tyagi S, Li SY, Mdluli K, Andries K, et al. Mutations in pepQ Confer Low-Level Resistance to Bedaquiline and Clofazimine in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2016; 60(8):4590–9. Epub 2016/05/18. https://doi.org/10.1128/AAC.00753-16 PMID: 27185800; PubMed Central PMCID: PMCPMC4958187.

56. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic acids research. 2014; 42(Web Server issue):W320–4. Epub 2014/04/23. https://doi.org/10.1093/nar/gku316 PMID: 24753421; PubMed Central PMCID: PMCPMC4086106.

Appendix J

Exploring Protein Supersecondary Structure Through Changes in Protein Folding, Stability, and Flexibility

Chapter 9

Exploring Protein Supersecondary Structure Through Changes in Protein Folding, Stability, and Flexibility

Douglas E. V. Pires, Carlos H. M. Rodrigues, Amanda T. S. Albanaz, Malancha Karmakar, Yoochan Myung, Joicymara Xavier, Eleni-Maria Michanetzi, Stephanie Portelli, and David B. Ascher

Abstract

The ability to predict how mutations affect protein structure, folding, and flexibility can elucidate the molecular mechanisms leading to disruption of supersecondary structures, the emergence of phenotypes, as well guiding rational protein engineering. The advent of fast and accurate computational tools has enabled us to comprehensively explore the landscape of mutation effects on protein structures, prioritizing mutations for rational experimental validation.

Here we describe the use of two complementary web-based in silico methods, DUET and DynaMut, developed to infer the effects of mutations on folding, stability, and flexibility and how they can be used to explore and interpret these effects on protein supersecondary structures.

Key words Missense mutations, Protein stability and folding, Machine learning, Normal mode analysis, Graph-based signatures, DUET, DynaMut

1 Introduction

Proteins are marginally stable, versatile macromolecules involved in a large variety of biochemical processes which are strictly linked and regulated by their native conformation. Mutations leading to changes in protein folding, stability, and conformation can have large phenotypic consequences, responsible for the development of many genetic disorders [1–14], including cancers, and even responsible for changes in drug susceptibility [15–27]. While these effects are commonly thought about in terms of reduced protein stability, mutations leading to increased stability and rigidification of the molecule can be equally deleterious. Maintaining, or enhancing, protein stability, and the identification of mutations that do not negatively affect protein stability, also remains one of the most difficult and important challenges in protein engineering.

Alexander E. Kister (ed.), *Protein Supersecondary Structures: Methods and Protocols*, Methods in Molecular Biology, vol. 1958, https://doi.org/10.1007/978-1-4939-9161-7_9,

While experimental validation of protein thermodynamic parameters remains a laborious task, the development of novel robust and scalable computational methods (Table 1) has allowed for the evaluation of the complete landscape of structural effects of mutations in a protein system and their effects on protein stability and flexibility within minutes, enabling rapid mutation prioritization.

Using the concept of graph-based signatures, we have developed robust methods for quantitatively analyzing effects of single missense mutations on protein stability, flexibility, and interactions [9, 28–37]. DUET [37] (http://biosig.unimelb.edu.au/duet) is a machine learning-based approach that integrates and optimizes two complementary methods in an optimized predictor (mCSM-Stability [36] and SDM [38]) using support vector machines. This method enables the accurate assessment of the effects of mutations on protein folding and stability. DynaMut [28] (http://biosig.unimelb.edu.au/dynamut) is a novel method that takes into account molecular motions and, by combining the graph-based signatures with coarse-grained normal mode analysis, generates a consensus prediction of effects of mutations on the protein conformational repertoire. These methods together compose a powerful platform that allows users to navigate the landscape of mutations effects on folding, stability, and flexibility.

2 Materials

DUET and DynaMut are structure-based methods for assessing effects of single-point missense mutations on protein stability/folding and protein flexibility/conformation, respectively. For both methods, users are required to provide:

1. Wild-type protein structure in PDB format: For both methods, a wild-type structure of the protein of interest in the Protein Data Bank [39] format (.pdb) must be provided to perform the predictions. This can be either (a) an experimentally solved structure, with previously solved structures available in the Protein Data Bank, or (b) a model, for instance, obtained via comparative homology modeling (*see* **Note 1** on how to deal with oligomeric structures). We have previously shown that using homology models built using templates down to 25% sequence identity does not significantly reduce predictive performance of either method (*see* **Note 2**). Users have the option to either upload the structure file or provide the PDB accession code when they wish to use an experimental structure previously deposited into the PDB (http://www.rcsb.org or http://www.ebi.ac.uk/pdbe/) (*see* **Note 3**).
2. Mutation information: The user also needs to supply information on the mutation or mutations they wish to analyze,

Table 1
List of freely available webservers and software for predicting effects of single-point mutations on protein folding, thermostability, and flexibility

	Method	Technique	Data set	Correlation	DOI	Publication year
Folding	mCSM-Stability	Structural signatures	ProTherm—351 mutations	0.73	https://doi.org/10.1093/bioinformatics/btt691	2014
	SDM2	Environment-specific substitution tables	ProTherm—351 mutations	0.61	https://doi.org/10.1093/nar/gkx439	2017
	DUET	Integrated approach	ProTherm—351 mutations	0.71	https://doi.org/10.1093/nar/gku411	2014
	Eris	Physical force field with atomic modeling	ProTherm—351 mutations	0.35	https://doi.org/10.1038/nmeth0607-466	2007
	I-Mutant 2.0	Neighboring residue composition	ProTherm—351 mutations	0.29	https://doi.org/10.1093/nar/gki375	2005
	Auto-Mute	Delaunay tessellation	ProTherm—351 mutations	0.46	https://doi.org/10.1155/2014/278385	2014
	CUPSAT	Atom potentials and torsion angle potentials	ProTherm—351 mutations	0.37	https://doi.org/10.1093/nar/gkl190	2006
	MAESTRO	Statistical scoring functions	ProTherm—351 mutations	0.70	https://doi.org/10.1186/s12859-015-0548-6	2015
	FoldX	Empirical full-atom force field	ProTherm—351 mutations	0.35	https://doi.org/10.1093/nar/gki387	2005
	PoPMuSiC	Statistical potentials and neural networks	ProTherm—351 mutations	0.67	https://doi.org/10.1186/1471-2105-12-151	2011
	NeEMO	Residue interaction networks	ProTherm—351 mutations	0.67	https://doi.org/10.1186/1471-2164-15-S4-S7	2014
Thermal stability	HoTMuSiC	Statistical potentials	ProTherm—1626 mutations	0.59	https://doi.org/10.1038/srep23257	2015
	FireProt	Structural and evolutionary information	ProTherm—1152 mutations	87% precision	https://doi.org/10.1093/nar/gkx285	2017
Flexibility	DynaMut	Structural signatures and NMA	ProTherm (2004)—351 mutations	0.69	https://doi.org/10.1093/nar/gky300	2018

including (1) the chain identifier (one-letter code of the chain, which corresponds to the 22nd column of the coordinate section in the PDB file where the mutation occurs) (*see* **Note 1**) and (2) the mutation code, which consists of the one-letter amino acid residue code of the wild-type residue, the residue number position as in the PDB file (columns 23–26 of the coordinate section), and the one-letter code of the mutated residue (e.g., R282W denotes a mutation from arginine to tryptophan at residue position 282).

3 Methods

3.1 Predicting and Analyzing Effects of Mutation on Protein Stability and Folding with DUET

1. DUET is freely available as a user-friendly web interface and is compatible with most operating systems and browsers. Open up the prediction server, http://biosig.unimelb.edu.au/duet/stability, on a web browser of your preference.
2. Provide the wild-type protein structure of interest by either uploading a PDB file or supplying a valid four-letter PDB accession code (Fig. 1a).
3. DUET offers users the option of two prediction modes, (a) assessing stability effects of a single mutation or (b) systematically evaluating all possible mutations at a given residue position. For a single mutation, users need to provide the mutation information and the mutation chain. For systematic evaluation, the one-letter code of the mutated residue is omitted.

3.2 DUET Prediction Output

1. If a single mutation is provided, after processing, the results page is shown (Fig. 1b), which includes information about the mutation and the predicted effects on stability for DUET and for the individual methods (mCSM-Stability and SDM). An interactive molecular visualization is also shown, allowing users to inspect the wild-type residue environment.
2. For systematic evaluation of a given residue, the predicted effects on protein stability for all 19 possible mutations are shown in tabular format (Fig. 1c).
3. Predicted effects are given as the change in Gibbs Free Energy, ΔΔG (kcal/mol), with negative values denoting destabilizing mutations and positive values, stabilizing ones. While users should interpret the values in the context of the protein system being studied, previous studies have used a rule of thumb that highly destabilizing/stabilizing mutations are those with a predicted $|\Delta\Delta G| > 1.0$ kcal/mol; and moderately destabilizing/stabilizing mutations are those with a predicted |ΔΔG| between 0.5 and 1.0. *See* **Notes 4** and **5** for further information on how to interpret results.

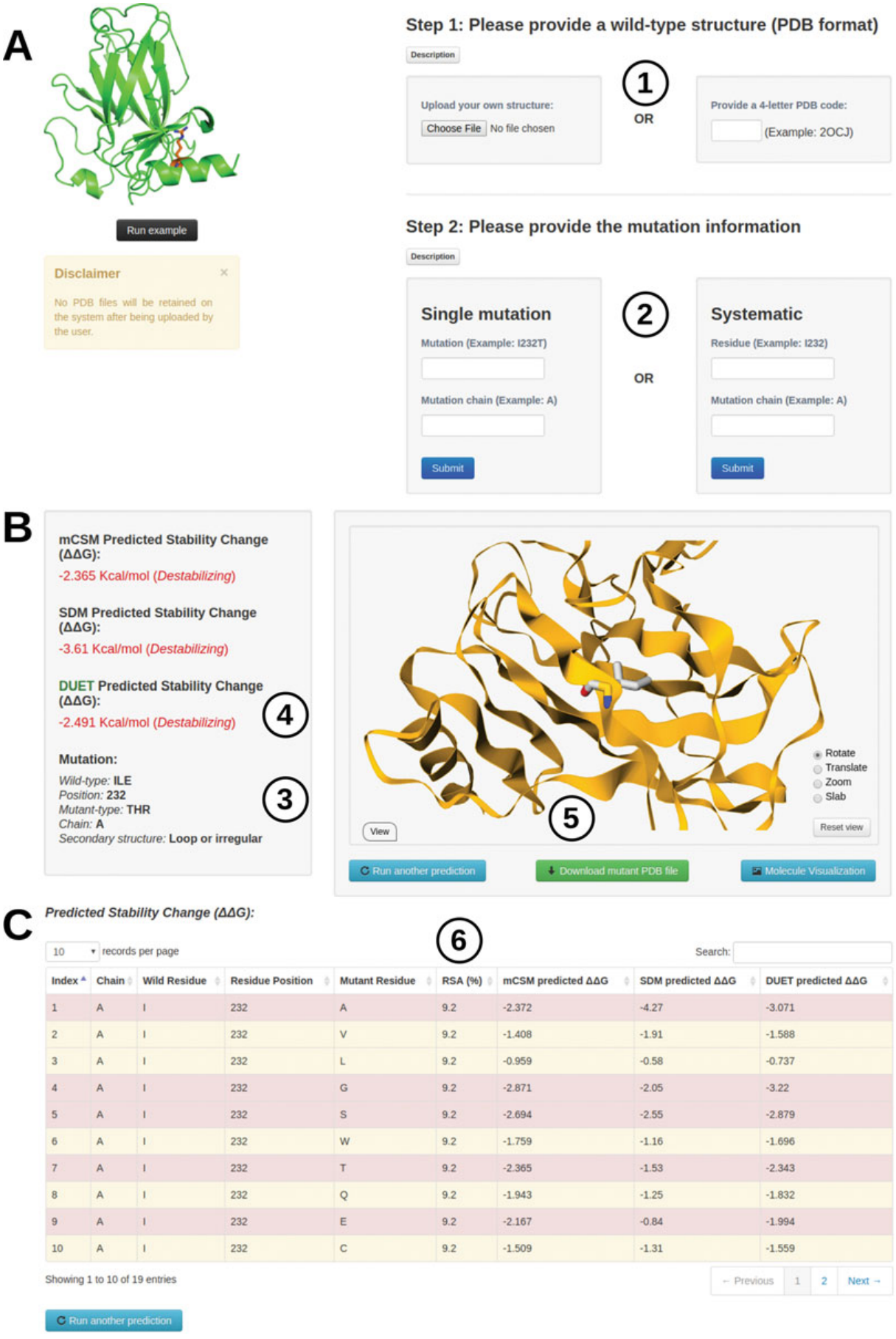


Fig. 1 DUET submission and results web interface. (**a**) The submission page allows users to either provide its own PDB file or inform an accession code of a protein of interest (*1*). Users have the option to analyze a

3.3 Predicting and Analyzing Effects of Mutations on Protein Flexibility and Conformation with DynaMut

1. As with DUET, DynaMut predicted changes upon mutation in protein stability are presented as a change in the Gibbs Free Energy of folding and stability ($\Delta\Delta G$ in kcal/mol), calculated as the difference between the wild-type and mutant proteins: $\Delta\Delta G = \Delta G_{wt} - \Delta G_{mt}$. A positive value denotes a stabilizing mutation, while a negative value denotes a destabilizing one. The DynaMut consensus prediction uses both normal mode analysis and graph-based signatures to more accurately identify stabilizing mutations, a limitation of other published approaches (Fig. 2b).
2. DynaMut is also freely available for use freely as a user-friendly web interface. In order to run a prediction, open up the DynaMut prediction page at http://biosig.unimelb.edu.au/dynamut/prediction on a web browser of your preference (the web server is compatible with the most common operating systems and browsers).
3. Users have the option to either evaluate a single mutation or provide a text file with a list of mutations to be evaluated in the same format discussed above to run DUET (Fig. 2a). There are no limits on the number of mutations that can be analyzed.
4. For both predictions modes, users are required to provide the wild-type protein structure of interest by either uploading a PDB file or supplying a valid four-letter code PDB accession code of a deposited experimental structure (Fig. 2a).

3.4 DynaMut Prediction Output

1. Prediction results: DynaMut will present the results under three main separate tabulated headings: (1) variation of Gibbs Free Energy predictions, (2) interatomic interactions, and (3) deformation/fluctuation analysis. *See* **Notes 4** and **5** for further information on how to interpret results.
2. DynaMut also graphically displays the resulting change in vibrational energy between the wild-type and mutant structures (Fig. 2b). This highlights regions predicted to be more flexible (red) or less flexible (blue) upon mutation. All calculations and representations can be downloaded through links located at the bottom of the results page.

Fig. 1 (continued) specific mutation or perform a systematic analysis of all mutations for a given residue (*2*). (**b**) For single-mutation prediction, the mutation identification (*3*) and the predicted effects on stability are shown (*4*), as well as an interactive molecular visualization (*5*). (**c**) For systematic evaluation of mutation on a given residue, the results are shown in tabular format

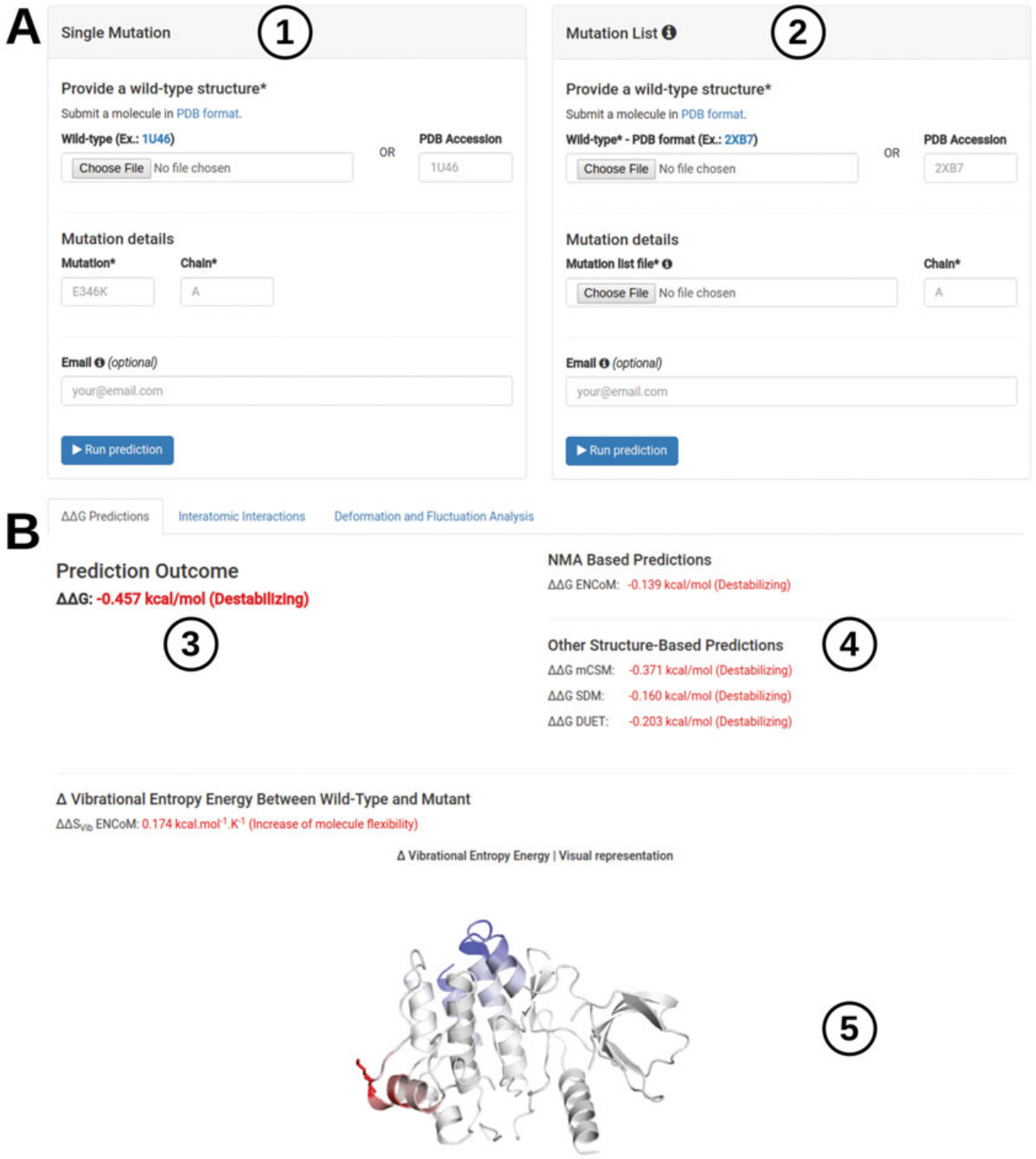


Fig. 2 DynaMut submission and results web interface. (**a**) The submission page allows for the analysis of a single-point mutation (*1*) or a list of mutations (*2*). The main results page (**b**) depicts the predicted effect of mutation by DynaMut (*3*) as well as predicted effects by its individual components (*4*). A depiction of the calculated different in vibration entropy (5) is also shown

3. When multiple mutations are analyzed, these results are presented in a tabulated format, where users are able to open up and analyze each mutation within the single-mutation analysis result interface.

3.5 Visualizing Effects of Mutations on Protein Structure

1. DynaMut also enables visualization of the effects of a mutation within the wild-type and mutant protein structure (Fig. 3).
2. The interatomic interactions made by the wild-type and mutant residues, calculated using Arpeggio [30] (http://biosig.unimelb.edu.au/arpeggioweb/), are visually shown. This enables the user to identify how the mutation will affect the local interaction network—important for maintaining protein stability (Fig. 3a).
3. The normal mode analysis predictions are also shown, highlighting changes in vibrational energy between the wild-type and mutant structures (Fig. 3b).
4. All these representations are downloadable as Pymol session files from links at the bottom of the results page.

4 Notes

1. It is important to notice that both methods, DUET and DynaMut, were conceived to analyze monomer structures. In case of analysis of oligomers, users are advised to filter their PDB files prior to submission, filtering chains of interest (for instance, using the PDBest software [40]). The servers will consider all chains submitted; however, a warning message is exhibited. When considering the effects of mutations on oligomeric structures, it is also important to consider the effects of the mutations on the affinity of the monomers to form the oligomer. This can be assessed using mCSM-PPI (http://biosig.unimelb.edu.au/mcsm/protein_protein).
2. The chain ID for the provided PDB file is a mandatory field, and blank characters are not allowed. Some homology modeling tools do not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. There are several tools available to perform this task.[1]
3. Another source of error comes from structures with multiple models. It is an important practice to filter NMR structures, selecting a single model.
4. Special cases: Mutations to and from prolines. Prolines are the only amino acid whose amino group is connected to the side chain, which in the context of the peptide bond greatly limits torsional angles. The nature of this residue, therefore, needs to be taken into account while analyzing mutation effects. For instance, (1) mutations to prolines in the middle of alpha-helices can introduce kinks, affecting local structure, and

[1] http://www.canoz.com/sdh/renamepdbchain.pl

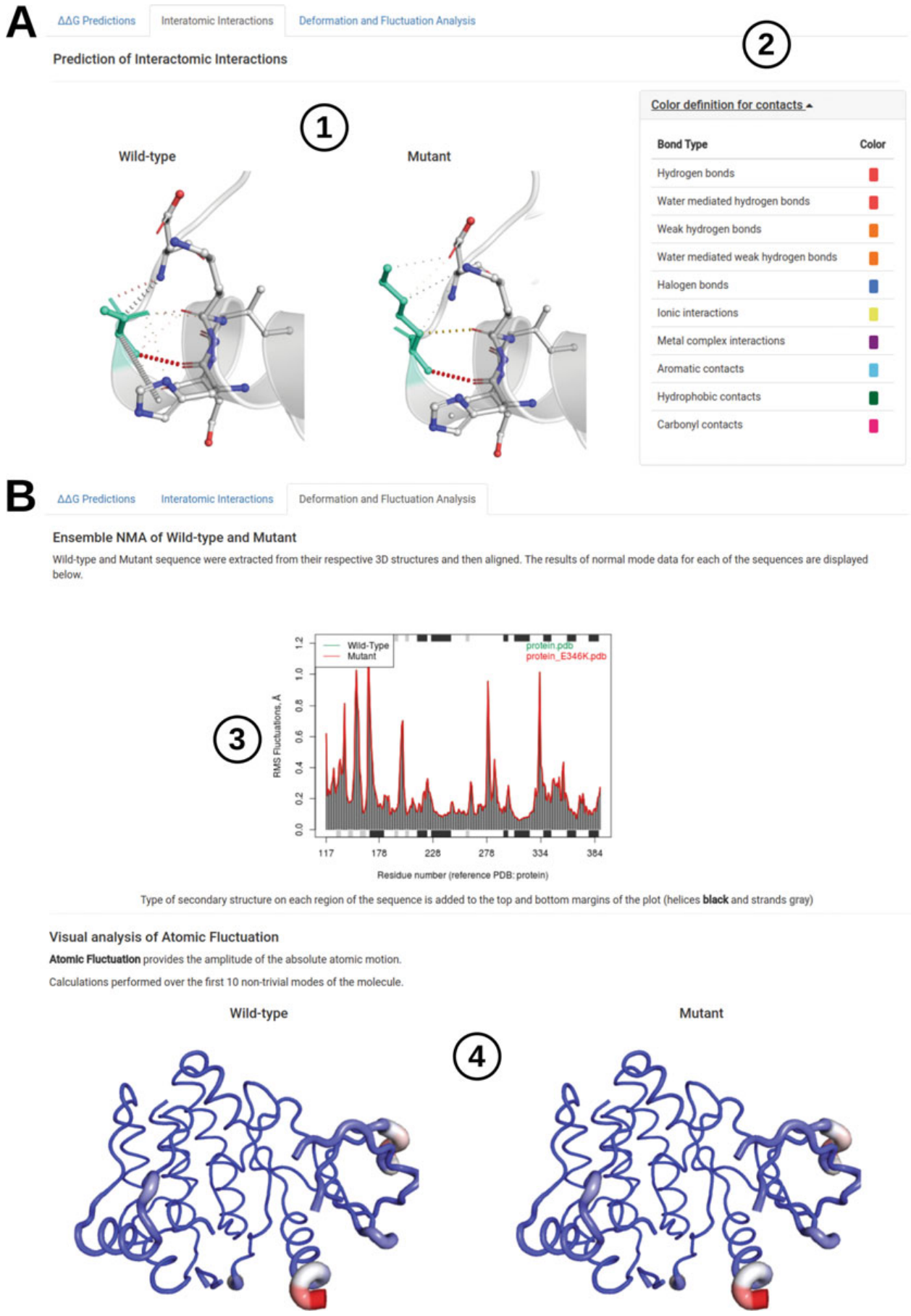


Fig. 3 DynaMut secondary results web interface. (**a**) A depiction of the calculated interatomic interactions (*1*) for wild-type and mutant proteins is shown, with interactions identified by color (*2*). (**b**) Depicts visualizations of the deformation and fluctuation analysis as fluctuation plot per residue (*3*) and atomic fluctuation in the context of the structures (*4*). Figure and individual files (pymol files for molecular visualization) are available for download

(2) since prolines are commonly found in turns and loops, their substitution might interfere with the formation of supersecondary structures such as hairpin loops.

5. Special cases: mutations of positive-phi glycines. Similarly to prolines, positive-phi glycines, while rare in experimental structures, should also be given special consideration due to its torsional angles. Glycines are the only residues capable of adopting positive-phi angles. These glycines are usually conserved across evolution, meaning that mutations of positive-phi glycines tend to be destabilizing.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarship [to Y.M., M.K., C.H.M.R. and S.P.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; the Victorian Life Sciences Computation Initiative (VLSCI), an initiative of the Victorian Government, Australia, on its Facility hosted at the University of Melbourne [UOM0017]; the Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P.]; and the Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.].

References

1. Andrews KA, Ascher DB, Pires DEV, Barnes DR, Vialard L, Casey RT, Bradshaw N, Adlard J, Aylwin S, Brennan P, Brewer C, Cole T, Cook JA, Davidson R, Donaldson A, Fryer A, Greenhalgh L, Hodgson SV, Irving R, Lalloo F, McConachie M, McConnell VPM, Morrison PJ, Murday V, Park SM, Simpson HL, Snape K, Stewart S, Tomkins SE, Wallis Y, Izatt L, Goudie D, Lindsay RS, Perry CG, Woodward ER, Antoniou AC, Maher ER (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55(6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
2. Trezza A, Bernini A, Langella A, Ascher DB, Pires DEV, Sodi A, Passerini I, Pelo E, Rizzo S, Niccolai N, Spiga O (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58 (12):5320–5328. https://doi.org/10.1167/iovs.17-22158
3. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27(10):1715–1729. https://doi.org/10.1101/gr.226589.117
4. Soardi FC, Machado-Silva A, Linhares ND, Zheng G, Qu Q, Pena HB, Martins TMM, Vieira HGS, Pereira NB, Melo-Minardi RC, Gomes CC, Gomez RS, Gomes DA, Pires DEV, Ascher DB, Yu H, Pena SDJ (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med

2:7. https://doi.org/10.1038/s41525-017-0009-4

5. Ramdzan YM, Trubetskov MM, Ormsby AR, Newcombe EA, Sui X, Tobin MJ, Bongiovanni MN, Gras SL, Dewson G, Miller JML, Finkbeiner S, Moily NS, Niclis J, Parish CL, Purcell AW, Baker MJ, Wilce JA, Waris S, Stojanovski D, Bocking T, Ang CS, Ascher DB, Reid GE, Hatters DM (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
6. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montano B, Blundell TL, Ascher DB (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
7. Chirgadze DY, Ascher DB, Blundell TL, Sibanda BL (2017) DNA-PKcs, allostery, and DNA double-strand break repair: defining the structure and setting the stage. Methods Enzymol 592:145–157. https://doi.org/10.1016/bs.mie.2017.04.001
8. Casey RT, Ascher DB, Rattenberry E, Izatt L, Andrews KA, Simpson HL, Challis B, Park SM, Bulusu VR, Lalloo F, Pires DEV, West H, Clark GR, Smith PS, Whitworth J, Papathomas TG, Taniere P, Savisaar R, Hurst LD, Woodward ER, Maher ER (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
9. Pires DE, Chen J, Blundell TL, Ascher DB (2016) In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
10. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenco CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
11. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
12. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, Rattenberry E, Simpson MA, Trembath RC, Weber A, Woodward ER, Donaldson A, Blundell TL, Latif F, Maher ER (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096
13. Hnizda A, Fabry M, Moriyama T, Pachl P, Kugler M, Brinsa V, Ascher DB, Carroll WL, Novak P, Zaliova M, Trka J, Rezacova P, Yang JJ, Veverka V (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia. https://doi.org/10.1038/s41375-018-0073-5
14. Sibanda BL, Chirgadze DY, Ascher DB, Blundell TL (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355 (6324):520–524. https://doi.org/10.1126/science.aak9654
15. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, Arumugam S, Rajan L, Ebenezer M, Ascher DB, Arnold E, Blundell TL (2018) Structural implications of mutations conferring rifampin resistance in mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1
16. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, Denholm JT, Ascher DB (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med. https://doi.org/10.1164/rccm.201712-2572LE
17. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTMH, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for EsxW Beijing variant in Vietnam. Nat Genet 50:849–856
18. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, Keri G, Ascher DB, Mondesert G, Vocat A, Lupien A, Sommer R, Vermet H, Lagrange S, Buechler J, Warner DF, McKinney JD, Pato J, Cole ST, Blundell TL, Rizzi M, Mizrahi V (2017) The inosine monophosphate dehydrogenase, GuaB2, is a

vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis 3(1):5–17. https://doi.org/10.1021/acsinfecdis.6b00102

19. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, Arora K, Ioerger TR, Sacchettini J, Rizzi M, Donini S, Blundell TL, Ascher DB, Rhee K, Breda A, Zhou N, Dartois V, Jonnala SR, Via LE, Mizrahi V, Epemolu O, Stojanovski L, Simeons F, Osuna-Cabello M, Ellis L, MacKenzie CJ, Smith AR, Davis SH, Murugesan D, Buchanan KI, Turner PA, Huggett M, Zuccotto F, Rebollo-Lopez MJ, Lafuente-Monasterio MJ, Sanz O, Diaz GS, Lelievre J, Ballell L, Selenski C, Axtman M, Ghidelli-Disse S, Pflaumer H, Bosche M, Drewes G, Freiberg GM, Kurnick MD, Srikumaran M, Kempf DJ, Green SR, Ray PC, Read K, Wyatt P, Barry CE 3rd, Boshoff HI (2017) Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against mycobacterium tuberculosis. ACS Infect Dis 3(1):18–33. https://doi.org/10.1021/acsinfecdis.6b00103
20. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
21. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
22. White RR, Ponsford AH, Weekes MP, Rodrigues RB, Ascher DB, Mol M, Selkirk ME, Gygi SP, Sanderson CM, Artavanis-Tsakonas K (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11):e1005977. https://doi.org/10.1371/journal.ppat.1005977
23. Silvino AC, Costa GL, Araujo FC, Ascher DB, Pires DE, Fontes CJ, Carvalho LH, Brito CF, Sousa TN (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172
24. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, Coeck N, Hill-Cawthorne GA, Nair MB, Mallard K, Ramsay A, Campino S, Hibberd ML, Pain A, Rigouts L, Clark TG (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14:31. https://doi.org/10.1186/s12916-016-0575-9
25. Kano FS, Souza-Silva FA, Torres LM, Lima BA, Sousa TN, Alves JR, Rocha RS, Fontes CJ, Sanchez BA, Adams JH, Brito CF, Pires DE, Ascher DB, Sell AM, Carvalho LH (2016) The presence, persistence and functional properties of Plasmodium vivax duffy binding protein II antibodies are influenced by HLA class II allelic variants. PLoS Negl Trop Dis 10(12):e0005177. https://doi.org/10.1371/journal.pntd.0005177
26. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
27. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom. https://doi.org/10.1099/mgen.0.000165
28. Rodrigues CHM, Pires DEV, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. https://doi.org/10.1093/nar/gky300
29. Pires DE, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45:W241–W246. https://doi.org/10.1093/nar/gkx236
30. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
31. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
32. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44(W1):W557–W561. https://doi.org/10.1093/nar/gkw390
33. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 44(W1):

W469–W473. https://doi.org/10.1093/nar/gkw458

34. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
35. Pires DE, Blundell TL, Ascher DB (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 58 (9):4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104
36. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
37. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42. (Web Server issue: W314–W319. https://doi.org/10.1093/nar/gku411
38. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45:W229–W235. https://doi.org/10.1093/nar/gkx439
39. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
40. Goncalves WR, Goncalves-Almeida VM, Arruda AL, Meira W Jr, da Silveira CH, Pires DE, de Melo-Minardi RC (2015) PDBest: a user-friendly platform for manipulating and enhancing protein structures. Bioinformatics 31(17):2894–2896. https://doi.org/10.1093/bioinformatics/btv223

Appendix K

mCSM-AB2: Guiding Rational Antibody Design Using Graph-Based Signatures

Bioinformatics, 36(5), 2020, 1453–1459
doi: 10.1093/bioinformatics/btz779
Advance Access Publication Date: 26 October 2019
Original Paper

Structural bioinformatics

mCSM-AB2: guiding rational antibody design using graph-based signatures

Yoochan Myung[1,2,3], Carlos H. M. Rodrigues[1,2,3], David B. Ascher[1,2,3,4,*] and Douglas E. V. Pires [1,2,3,5,*]

[1]Department of Biochemistry and Molecular Biology, [2]ACRF Facility for Innovative Cancer Drug Discovery, Bio21 Institute, University of Melbourne, Melbourne, VIC 3010, Australia, [3]Structural Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, VIC 3004, Australia, [4]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK and [5]School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010, Australia

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on May 9, 2019; revised on October 7, 2019; editorial decision on October 8, 2019; accepted on October 23, 2019

Abstract

Motivation: A lack of accurate computational tools to guide rational mutagenesis has made affinity maturation a recurrent challenge in antibody (Ab) development. We previously showed that graph-based signatures can be used to predict the effects of mutations on Ab binding affinity.

Results: Here we present an updated and refined version of this approach, mCSM-AB2, capable of accurately modelling the effects of mutations on Ab–antigen binding affinity, through the inclusion of evolutionary and energetic terms. Using a new and expanded database of over 1800 mutations with experimental binding measurements and structural information, mCSM-AB2 achieved a Pearson's correlation of 0.73 and 0.77 across training and blind tests, respectively, outperforming available methods currently used for rational Ab engineering.

Availability and implementation: mCSM-AB2 is available as a user-friendly and freely accessible web server providing rapid analysis of both individual mutations or the entire binding interface to guide rational antibody affinity maturation at http://biosig.unimelb.edu.au/mcsm_ab2

Contact: david.ascher@unimelb.edu.au or douglas.pires@unimelb.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Antibodies (Abs) are central components of our immune system that bind specifically to their target antigens in order to elicit an immune response. This interaction between an Ab and its antigen is mediated by a myriad of non-covalent interactions made by the complementary determining regions (CDRs) of the Abs with a specific epitope on an antigen. This ability to bind to a wide variety of targets, including those traditionally considered undruggable, in a highly specific and selective manner has led to increasing interest in their use as therapeutics for a broad range of diseases including several types of cancer (Elgundi *et al.*, 2017) and rheumatoid arthritis (Tanaka *et al.*, 2014). Since the first approval of monoclonal Ab, the significant improvement in Ab engineering has led Abs to become best-selling drugs accounting for over half of the therapeutic market (Urquhart, 2018).

Ab development often requires optimization of its stability, solubility, selectivity, affinity and immunogenicity. Achieving the desired properties can often become a major challenge, considering the large number of possible variations in Abs, and with each potentially affecting multiple biological properties. One of the early steps in the development of effective Ab therapies is the engineering of binding specificities and selectivities, which has traditionally been inspired by the natural biological process of affinity maturation, with rounds of mutations within the CDR loops explored. This process can be time-consuming, and is inherently a random and error-prone process. Recent examples, however, have shown how the computationally guided rational engineering of Ab-binding affinities can dramatically improve this process (Kiyoshi *et al.*, 2014; Sefid *et al.*, 2019).

A number of different computational approaches which use an available crystal structure to guide Ab design and optimization have been developed (Roy *et al.*, 2017). A systematic evaluation of the accuracy of these approaches to predict the change upon mutation in binding affinity highlighted the limited performance of existing tools, and the challenging nature of this problem.

In a previous work, we have adapted the concept of graph-based signatures which can efficiently represent the physicochemical properties and geometry of surrounding environment of the wild-type and mutant residues to accurately predict the effects of mutations in terms of protein stability (Pandurangan *et al.*, 2017a; Pires *et al.*, 2014a, b;

Rodrigues *et al.*, 2018b) and interactions with other proteins (Pires *et al.*, 2014b; Rodrigues *et al.*, 2019), nucleic acids (Pires *et al.*, 2014b; Pires and Ascher, 2017), small molecules (Pires *et al.*, 2015; Pires and Ascher, 2016a) and metal ions (Pires *et al.*, 2016b). These have been successfully used to provide valuable insights into genetic diseases (Albanaz *et al.*, 2017; Andrews *et al.*, 2018; Ascher *et al.*, 2019; Casey *et al.*, 2017; Hnizda *et al.*, 2018; Jafri *et al.*, 2015; Jubb *et al.*, 2017; Nemethova *et al.*, 2016; Pandurangan *et al.*, 2017b; Ramdzan *et al.*, 2017; Rodrigues *et al.*, 2018a; Silvino *et al.*, 2016; Soardi *et al.*, 2017; Traynelis *et al.*, 2017; Trezza *et al.*, 2017; Usher *et al.*, 2015), drug resistance (Ascher *et al.*, 2015; Hawkey *et al.*, 2018; Holt *et al.*, 2018; Karmakar *et al.*, 2018, 2019; Phelan *et al.*, 2016; Pires *et al.*, 2016a, b; Portelli *et al.*, 2018; Vedithi *et al.*, 2018) and rational protein engineering. We have also successfully applied our graph-based signatures to the prediction of changes in Ab–antigen binding affinity and showed that this outperformed existing methods, although there was still significant room for improvement (Pires and Ascher, 2016b). The release of SKEMPI2.0 containing information of the effects of new mutations on Ab–antigen binding affinity, allowed us to not only assess earlier approaches based on new unseen experimental data, but to also build a predictive model across a more comprehensive set of Ab–antigen complexes and mutations. In particular, mCSM-AB only considered structural information, however evolutionary information and energetic terms have been shown to help predict the effect of a mutation on Ab-binding affinity, as variants which have destabilizing effects on proteins are less likely to be conserved from an evolutionary perspective (Gonzalez-Munoz *et al.*, 2012). In addition, Ab–antigen interfaces are enriched with specific type of amino acids such as Tyr and Ser (Jubb *et al.*, 2015; Van Regenmortel, 2014) compared with other protein–protein complexes, and different modes of interatomic interaction may be important to explain whether the mutation is favourable in its surroundings.

A powerful and scalable model for predicting the effects of missense mutations on Ab-binding affinity could hold enormous potential for guiding rational Ab development. Here we introduce mCSM-AB2, an updated and optimized version of our previous method, trained on a larger and more comprehensive dataset, which uses not only graph-based signatures but also interatomic interaction, evolutionary and energy-based features to capture additional structural and sequence-based information to more accurately predict Ab–antigen affinity changes upon mutation. We show that mCSM-AB2 significantly outperforms existing methods, and has potential to guide rational Ab engineering.

2 Materials and methods

The general mCSM-AB2 workflow is depicted in Figure 1. It is composed of three main steps including: (i) dataset acquisition, which refers to collecting experimental evidence from the literature on effects of mutations in Ab–antigen binding affinity complexes with solved structures; (ii) feature engineering, which encompasses the generation and evaluation of features selected to model different aspects involved in Ab–antigen recognition and effects of mutations

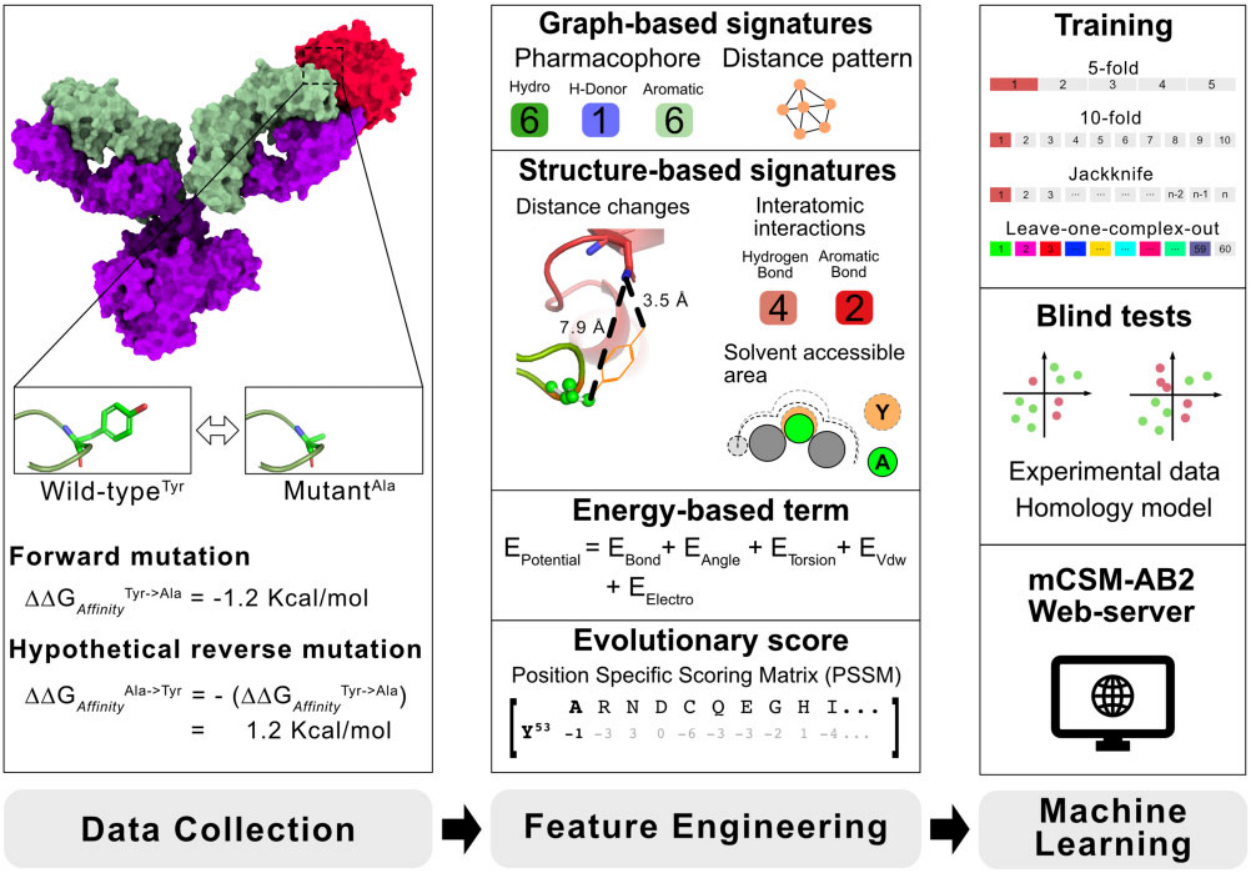


Fig. 1. Overview of the mCSM-AB2 workflow. In data collection, after data acquisition, hypothetical reverse mutations are considered to avoid the natural bias of mutations reducing affinity. From the complete dataset (1810 mutations), a range of different features are calculated to be used as evidence to train predictive models using machine learning algorithms. Among the feature classes, graph-based signatures are used to describe the wild-type residue environment and its geometry and physicochemical properties. Other structural attributes aiming to model other relevant aspects driving Ab-antigen affinity were also considered, including the variation in the distance to the antigen upon mutations, solvent accessible area, as well as energetic terms and interatomic interactions. Additionally, an evolutionary score (derived from PSSM) was also used to model mutation tolerance throughout evolution. All features are then used to build predictive models through a series of training and blind-test validation procedures. The best model was then made available as an easy-to-use web server at http://biosig.unimelb.edu.au/mcsm_ab2

on these complexes and (iii) machine learning, which aims to train, test and validate an accurate predictive model via supervised learning, using the computed features and experimental effects of mutations, as evidence.

2.1 Datasets

To develop our predictive model, we collected binding affinity data with experimentally determined structures from the AB-BIND (Sirin *et al.*, 2016), PROXiMATE (Yugandhar *et al.*, 2017) and SKEMPI2.0 (Jankauskaite *et al.*, 2018) databases to train and test mCSM-AB2. Compared with the earlier AB-BIND dataset used for mCSM-AB, we discarded 3 redundant mutations and 27 mutant non-binders, which led to a dataset of 558 mutations that was used as the training set. SKEMPI2.0 contained 830 single point mutations from Ab–antigen complexes, which were filtered using 'Hold_out_type' and 'Hold_out_proteins' information, to avoid redundancy during training. Of these 830 instances, there were 102 mutations that had more than one experimentally measured binding affinity for the same mutation, for which we preferentially kept direct binding assays such as SPR or ITC, leading to a group of 728 mutations. Within the 728 mutations selected from SKEMPI2.0, we also disregarded 32 mutant non-binders, leaving a total of 696 mutations in Ab–antigen complexes. Comparing to AB-BIND, our filtered SKEMPI2.0 dataset contained 377 new mutations in unique complexes not present in the original training set. CD-HIT (Huang *et al.*, 2010) was used to cluster interfaces with a similarity threshold of 90%. The new SKEMPI2.0 dataset contained unique interfaces not present in the original AB-BLIND dataset, and was hence used as a non-redundant blind test and comparative tool to evaluate previous methods built using AB-BIND. In total, our train/blind-test datasets are composed of 905 single mutations and the mutations from 11 out of 60 Ab–antigen complexes are present in both training (AB-BIND) and blind test sets (Supplementary Table S1). This represents not only a significant increase over the 558 mutations used to train mCSM-AB spanning across more than twice the number of 25 Ab-antigen complexes, but also a non-redundant experimental blind test set (377 mutations) which allows us to explicitly compare the performance of our new approach to mCSM-AB and other methods.

Due to the nature of experimental affinity maturation, these datasets were unbalanced, with 652 destabilizing ($\Delta\Delta G_{\text{Affinity}} < 0$), 196 stabilizing ($\Delta\Delta G_{\text{Affinity}} > 0$) and 57 neutral ($\Delta\Delta G_{\text{Affinity}} = 0$) single point mutations (Supplementary Figure S1, left). As has been previously proposed (Thiltgen and Goldstein, 2012), to avoid any subsequent bias in our predictive model we also considered the hypothetical reverse mutations, using mutant structures generated by FoldX (Eswar *et al.*, 2006). This gave a final dataset of 1810 single point mutations (Supplementary Figure S1, right), of which 1056 were used for training, and a non-redundant set of 754 mutations were used as a blind test set to avoid overtraining and to benchmark the performance of mCSM-AB2. We also used an additional blind test set of 87 mutations across five homology models as proposed previously (Pires and Ascher, 2016b). The datasets used to train and validate mCSM-AB2 are available on the mCSM-AB2 web server.

2.2 Feature engineering

Three main classes of features were used in mCSM-AB2 as evidence to train and test predictive models via supervised learning—structural, evolutionary and energy-based terms. Graph-based signature are calculated to model the wild-type residue environment.

These represent distance patterns between different atom types as cumulative distributions of distances, which we previously show encode both its physicochemical aspects and geometry (Pires *et al.*, 2014a, 2016a; Pires and Ascher, 2016a, 2017; Rodrigues *et al.*, 2018a). In order to calculate structure-based features for mutants, we implemented BuildModel of FoldX for high quality models. Additionally, the changes in pharmacophores due to the mutation are also modelled as a feature vector. These pharmacophore changes calculated the difference in atom counts per class (hydrophobic, positive charge, negative charge, hydrogen acceptor, hydrogen donor, aromatic, sulphur and neutral) between wild-type and mutant residues. Additional structural information was also taken into account, including the change in molecular interactions upon mutation as calculated by Arpeggio (Jubb *et al.*, 2017), the distance change of the mutation to the Ab–antigen interface, and the change of relative solvent accessible (RSA) area upon mutation using DSSP (Touw *et al.*, 2015). Evolutionary-based information was integrated by calculating the difference of evolutionary scores between wild-type and mutant using PAM30-based position-specific scoring matrices (PSSM) (Altschul *et al.*, 1997). An energy-based term was also generated using FoldX (Stricher *et al.*, 2005) force fields to calculate the difference upon mutation in potential energy between the wild-type and mutant structures, expressed in kcal/mol.

2.3 Machine learning methods

Using the collected experimental data describing the effects of missense mutations on Ab–antigen affinity and calculated features, different supervised learning algorithms available on the Scikit-learn library for Python (Pedregosa *et al.*, 2011) were evaluated, including Extra Trees (Geurts *et al.*, 2006), Random Forest (Breiman, 2001), Gradient Boost (Friedman, 2002) and XGBoost (Chen and Guestrin, 2016) regression. Predictive models were trained using five times stratified 10-fold cross-validation to avoid sampling bias, followed by a blind test. A leave-one-complex out cross-validation procedure was also implemented to assess performance variations for different Ab–antigen complexes. The final model showed comparable performances across the different training schemes including 5-fold, 10-fold, leave-one-complex-out and Jackknife (Wager *et al.*, 2014) validation, as shown in Supplementary Table S2.

2.4 Evaluation metrics

The performance of individual models was assessed using the Pearson's correlation coefficient and root mean square error (RMSE), considering performances on both cross-validation and blind tests. The performance of the model was also assessed on 90% of the data after removing 10% of worst predicted cases to evaluate effects of outliers on model accuracy.

3 Results

In order to evaluate the performance of mCSM-AB2, we devised a series of experiments. The first aim was to assess the contribution of individual feature components to predictive performance as well as their combination. mCSM-AB2 was further tested on blind tests and its performance was compared with available methods.

3.1 Quantitative assessment of Ab–antigen affinity changes upon mutation

Building upon the previous version of mCSM-AB, we have integrated new structure-based features, energy-based terms and evolutionary scores with our graph-based signatures to better model the changes of topological and physicochemical properties on Ab–antigen affinity induced by missense mutations. Supplementary Table S2 shows the predictive performance of the individual feature classes, given as Pearson's correlation coefficient, for different validation procedures, including 5- and 10-fold cross-validation, as well as Jackknife validation.

The best performing individual class of features was the graph-based signatures, contributing to a correlation of $\rho = 0.65$ (RMSE of 2.14 kcal/mol) on 10-fold cross-validation, followed by the difference in contacts made by wild-type and mutant residues, which achieved a correlation of $\rho = 0.60$ (RMSE of 2.40 kcal/mol), highlighting the important role of inter-residue interactions on driving Ab–antigen affinity and recognition. Pharmacophore modelling was also an important feature class, achieving a correlation of $\rho = 0.50$ (RMSE of 3.12 kcal/mol). Complementary structure-based information was also integrated to the method, even with modest performance, including the change of the RSA upon mutation ($\rho = 0.16$

and RMSE of 10.92 kcal/mol), the change of distance from mutation site to the antigen interface ($\rho = 0.26$ and RMSE of 6.64 kcal/mol).

Other two features incorporated on this new and updated version of the method were energy potential terms calculated using FoldX and sequence-based evolutionary information encoded in PSSM scoring matrices. These features contributed individually to a predictive performance of $\rho = 0.26$ (RMSE of 6.61 kcal/mol) and $\rho = 0.42$ (RMSE of 3.95 kcal/mol), respectively.

It is interesting to notice that there seems to be little correlation between the different classes of selected features, as shown in Supplementary Figure S2, especially to the new evolutionary and energy-based attributes, indicating they were likely contributing to the predictive model with non-redundant, novel information. In addition, regardless of lower performance of evolutionary- and energy-based features, those have greater importance on the mCSM-AB2 model which indicates those two features high chance to give synergistic effect with other features, not by themselves (Supplementary Figure S3).

By combining the different feature classes to train a regressor algorithm/model, we obtained an improved and optimized model capable of accurately and quantitatively predicting effects of mutations on Ab–antigen binding affinity across eight different algorithms, achieving a Pearson's correlation coefficient of $\rho = 0.73$ (RSME of 1.68 kcal/mol) from Extra Tress algorithm (Supplementary Fig. 2A, Table S3) on 10-fold cross-validation. This model was significantly different ($P \ll 0.05$ by Diebold–Mariano test) compared with the null hypothesis using the average of all values as the prediction (RMSE = 1.80 kcal/mol), the average of just the experimentally measured changes in binding affinity (RMSE = 2.07 kcal/mol), and by randomly scrambling the $\Delta\Delta G$ 10 times to keep the same data distribution (RMSE = 2.56 kcal/mol). The performance of the method increases to $\rho = 0.84$ on 90% of the data and was not significantly different when either 5-fold cross-validation or Jackknife validation were used, providing additional confidence in the model.

Compared with earlier mCSM-AB, we implemented additional features from both wild-type and mutant structures which demand more computational cost, but those features improved the performance on training and two blind tests (Supplementary Table S4). The reliability of the model structures obtained through FoldX was assessed by comparing with seven experimental mutant structures (Supplementary Table S5). The modelled structures used in mCSM-AB2 showed a low average C_α RMSD of 0.13 Å.

3.2 Comparative performance and blind tests

In order to put mCSM-AB2 prediction results into context, we have carried out a performance comparison with other available methods using a non-redundant blind test composed of 754 mutations with experimentally measured changes in binding affinity. mCSM-AB2 significantly outperformed alternative approaches, achieving a Pearson's correlation coefficient of $\rho = 0.64$ ($P \leq 0.0001$, as depicted in Table 1 and Fig. 2B), showing that not only it was able to accurately predict Ab–antigen binding affinity changes but also presented a significant improvement in comparison with its previous version ($\rho = 0.42$). This performance was comparable to the cross-validation performance, increasing our confidence in the method's generalization capabilities.

Comparison of mCSM-AB2 performance across the training set also showed it performed significantly better than other methods that have been used to guide rational Ab engineering (Table 1). Interestingly, there were only weak correlations between mCSM-AB2 and other Ab engineering methods (Supplementary Fig. S4), including the original method, highlighting its use of complementary but distinguishing information, and suggesting that a consensus predictor might be informative.

The experimental datasets were enriched in mutations located at the antigen interface (>80% within 6 Å as shown in Fig. 2C), which is not surprising since many experiments have focused on variations in the CDR loops with alanine scanning (>60% of mutations in the dataset are to alanine). The distance from a mutation site to the Ab–antigen interface influenced on the performance of mCSM-AB2. Comparing performance on mutations less than 6 Å, 6–10 Å and greater than 10 Å away from the antigen interface, mCSM-AB2 achieved a Pearson's correlation of 0.74 ($\sigma = 0.004$), 0.52 ($\sigma = 0.029$) and 0.54 ($\sigma = 0.073$), respectively. This deterioration of the performance on mutations located further away from the interface may be due to the limited number of distal mutations in the training set. As a result of the distance-based analysis, the mCSM-AB2 web server gives users a confidence level of prediction, high or moderate, depending on the distance between the mutation and Ab–antigen binding interface.

Table 1. Performance comparison between mCSM-AB2 and available methods

Method	Pearson's correlation		
	Training	Blind test	
	AB-BIND	Experimental set	Homology model
bASA	0.22[a],***	0.29***	0.41***
dDFIRE	0.19[a],***	0.31***	0.53***
DFIRE	0.31[a],***	0.38***	0.52**
FoldX	0.34[a],***	0.26***	0.45***
Discovery Studio	0.45[a],***	0.31***	0.53**
mCSM-PPI	0.35[a],***	0.32***	0.26***
mCSM-AB	0.56[a],***	0.42***	0.54*
mCSM-AB2	0.76	0.64	0.77

Note: Pearson's correlation coefficient of each of the methods were compared with mCSM-AB2 by Fisher's *r*-to-*z* transformation (*$P \leq 0.05$, **$P \leq 0.001$ and ***$P \leq 0.0001$).

[a]From Sirin *et al.* (2016).

While mutations to alanine were inherently enriched in the dataset, the performance of mCSM-AB2 was consistent across mutations to any residue (Supplementary Table S6). This can be further supported by the analysis of the experimental blind test results showing mCSM-AB2 outperforms all other methods across all types of mutations (Supplementary Fig. S5).

An earlier study (Sinha *et al.*, 2002) suggested several experimental $\Delta\Delta G$s from the HyHEL-10 Fab and lysozyme complex (PDB: 3HFM), which were measured by indirect methods such as spectroscopic inhibition assay (IASP) and spectroscopic method (SP), presented a large discrepancy with $\Delta\Delta G$ from direct method such as surface plasmon resonance (SPR). In order to measure the contribution of each of Ab–antigen complexes on the performance of mCSM-AB2, we conducted the leave-one-complex-out cross-validation on the 60 Ab–antigen complexes. Notably, the mutations from 3HFM presented a large portion of outliers in both 10-fold and leave-one-complex-out cross validations showing 31 and 17 out of 181 worst predicted data points, respectively (Supplementary Table S7). The overall performance on leave-one-complex-out (Pearson's correlation of $\rho = 0.70$), however, was comparable with the 10-fold cross-validation results (Pearson"s correlation of $\rho = 0.73$), further demonstrating the robustness of the method.

3.3 Performance on homology models

As experimental crystal structures might not always be available, we also wanted to compare the performance of mCSM-AB2 on predicting effects of mutations on Ab–antigen binding affinity using homology models. We used a previously proposed homology model dataset (Sirin *et al.*, 2016) of 87 experimentally measured changes in binding affinity upon mutation across five homology models of the corresponding Ab–antigen complex. The mCSM-AB2 predictions correlated well with the experimental values ($\rho = 0.77$, RMSE = 1.66), and was significantly more accurate than all other predictive methods analyzed (Table 1). This highlights the versatility and robustness of the mCSM-AB2 predictions, and its applicability even in the absence of an experimental structure of an Ab–antigen complex.

3.4 mCSM-AB2 web server

We have developed a web server to provide the functionalities of mCSM-AB2 in an intuitive way, increasing reproducibility and

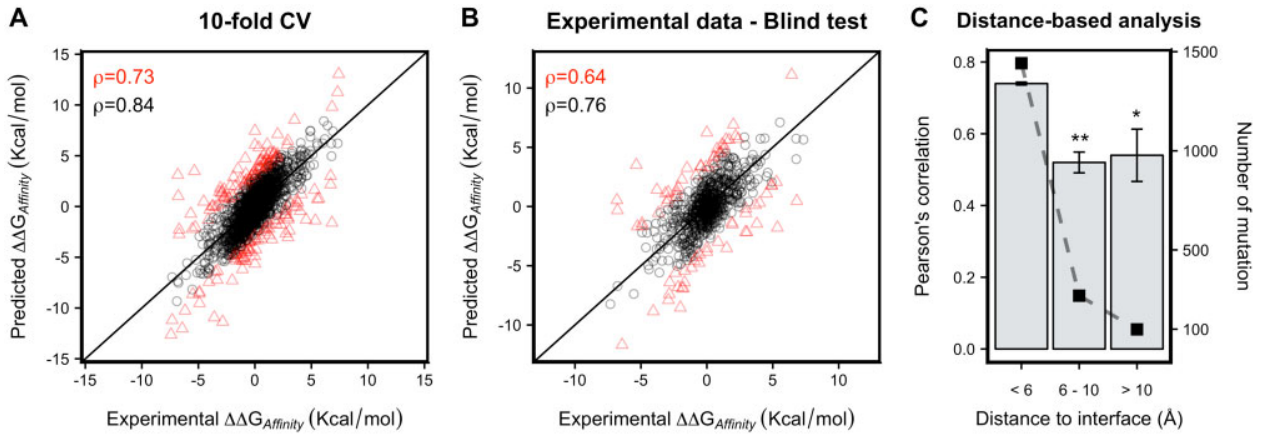


Fig. 2. Performance of mCSM-AB2 in predicting Ab–antigen affinity changes upon mutation. mCSM-AB2 achieved a Pearson's correlation of $\rho = 0.73$ (RMSE = 1.68) on 10-fold cross-validation (**A**), $\rho = 0.64$ (RMSE = 1.85) on a non-redundant experimental dataset for a model trained on the AB-BIND dataset (**B**). Performance of mCSM-AB2 after excluding the 10% largest errors (red triangles) are shown as black circles. (**C**) Through 10 times of 10-fold cross-validation runs, mCSM-AB2 achieved a Pearson's correlation of 0.74 ($\sigma = 0.004$), 0.52 ($\sigma = 0.029$) and 0.54 ($\sigma = 0.073$) on mutations whose distances to their Ab–antigen binding interfaces are less than 6 Å (1442 mutations), between 6 and 10 Å (269 mutations) and greater than 10 Å (99 mutations), respectively. Fisher's *r*-to-*z* transformation was used to compare Pearson's correlations from different size of mutation (* $P \leq 0.001$ and ** $P \leq 0.0001$). (Color version of this figure is available at *Bioinformatics* online.)

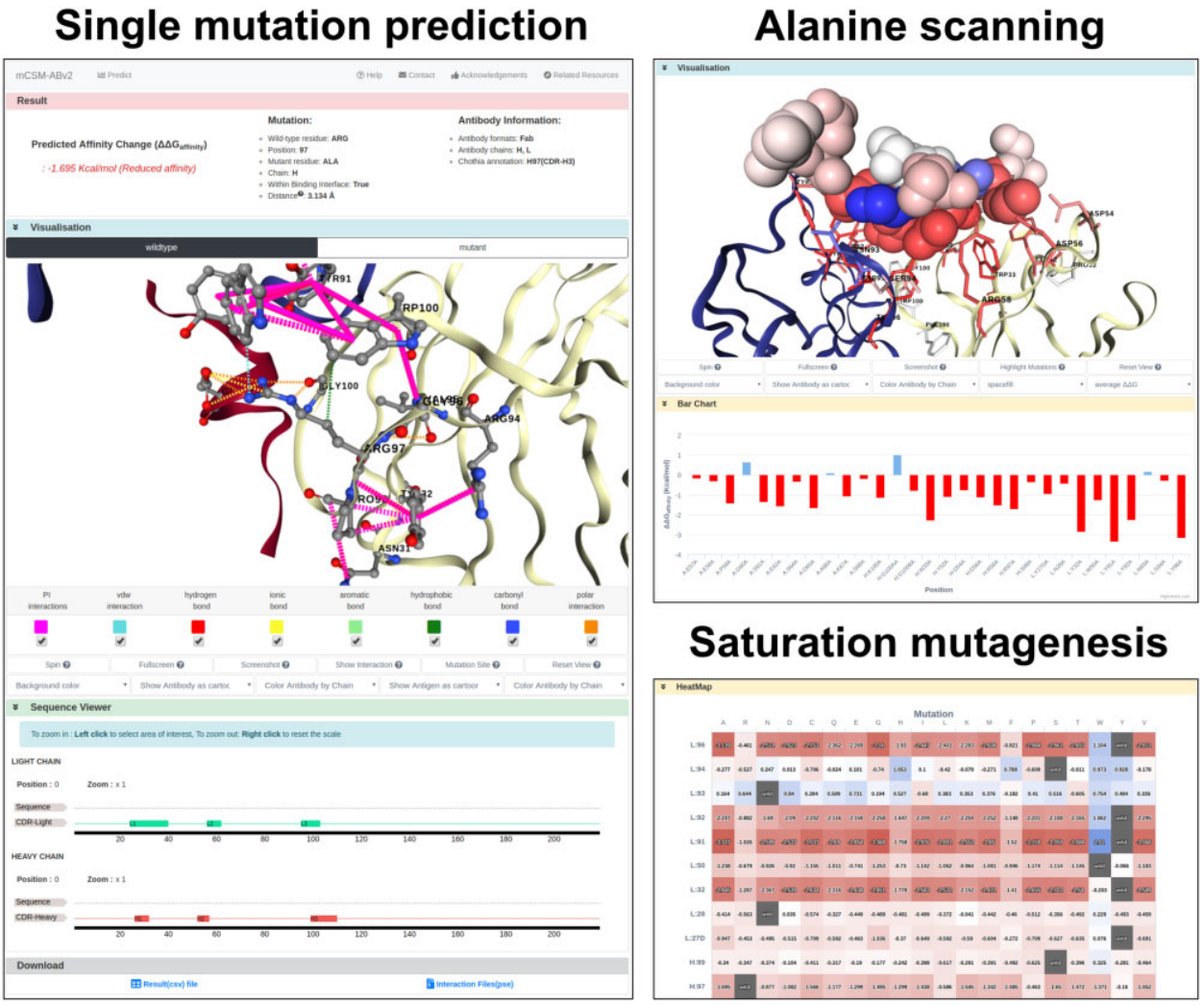


Fig. 3. mCSM-AB2 web server result pages. Single mutation prediction (left) provides predicted $\Delta\Delta G_{\text{Affinity}}$ and interaction changes upon mutation via a 3D molecular viewer for both wild-type and mutant. Alanine scanning (right top) describes mutational effects on interface residues with molecular viewer and bar charts. In saturation mutagenesis analysis (right bottom), users can check Ab–antigen affinity changes for each of the 19 possible mutations for each interface residues

facilitating large-scale analyses. The front-end was designed with Bootstrap framework version 4.1 and the back-end was based on Python 2.7 via the Flask framework version 1.0.2 on a Linux server running Apache. It allows users to upload Ab–antigen complexes (in PDB format) and either analyze specific mutations provided by the user, or systematically evaluate mutations across the entire Ab–antigen interface via either alanine scanning or saturation mutagenesis, facilitating, for instance, the identification of mutations that are more likely to increase affinity, aiding the rational design of Abs. The results pages allow easy visualization of the alanine-scanning and saturation-mutagenesis predictions mapped to the 3D structure as well as heat-mapped tables (Fig. 3). Users are able to check information such as distance to binding interface and Chothia annotation calculated by ANARCI (Dunbar and Deane, 2016) and download all results including the predictions as a CSV file, and the provided PDB files with the predicted changes in binding affinity mapped to the B-factor column.

4 Conclusions

The ability to predict favourable Ab–antigen mutations is a crucial, but non-trivial, challenge to help guide routine affinity maturation. While a number of successful computational-guided Ab development examples have been published in recent years, computational tools haven't had yet transformative effects for Ab engineering due to limited accuracy of available computational methods.

mCSM-AB2 is a computational approach that leverages both sequence and structural information to allow users to accurately assess the effects of single-point mutations on Ab–antigen binding affinity. Across all training and blind test evaluations, mCSM-AB2 significantly outperformed all currently used Ab mutational analysis approaches, using both experimental structures and homology models, highlighting its potential power to help guide Ab development. This also highlights the power of our graph-based signatures in terms of predicting mutational effects on Ab–antigen affinities by efficiently representing structural environment of wild-type and mutant residues, but also show the importance of considering evolutionary aspects, energetic terms and inter-residue interactions to better understand molecular recognition.

We believe that mCSM-AB2 will be a powerful tool to not only streamline Ab development and engineering but also providing better insight into the effects of mutations in Ab–antigen interfaces, including escape mutations. A user-friendly web server implementing mCSM-AB2 functionalities was implemented and is freely available at http://biosig.unimelb.edu.au/mcsm_ab2, facilitating large-scale analysis of entire Ab–antigen interfaces.

Funding

Y.M. and C.H.M.R were funded by the Melbourne Research Scholarship. D.B.A and D.E.V.P were funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); the Jack Brockhoff Foundation [JBF 4186, 2016]; and a C. J. Martin Research Fellowship from the National Health and Medical Research Council (NHMRC) of Australia [APP1072476]. Supported in part by the Victorian Government's OIS Program.

Conflict of Interest: none declared.

References

Albanaz,A.T.S. *et al.* (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin. Drug Discov.*, **12**, 553–563.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrews,K.A. *et al.* (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J. Med. Genet.*, **55**, 384–394.

Ascher,D.B. *et al.* (2015) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci. Rep.*, **4**, 4765.

Ascher,D.B. *et al.* (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.*, **27**, 888–902.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Casey,R.T. *et al.* (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol. Genet. Genomic Med.*, **5**, 237–250.

Chen,T. and Guestrin,C. (2016) XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, San Francisco*, pp. 785–794. ACM, San Francisco, CA.

Dunbar,J. and Deane,C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298–300.

Elgundi,Z. *et al.* (2017) The state-of-play and future of antibody therapeutics. *Adv. Drug Deliv. Rev.*, **122**, 2–19.

Eswar,N. *et al.* (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*, **Chapter 5**, Unit-5.6.

Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Statist. Data Anal.*, **38**, 367–378.

Geurts,P. *et al.* (2006) Extremely randomized trees. *Mach. Learn.*, **63**, 3–42.

Gonzalez-Munoz,A. *et al.* (2012) Tailored amino acid diversity for the evolution of antibody affinity. *MAbs*, **4**, 664–672.

Hawkey,J. *et al.* (2018) Evolution of carbapenem resistance in *Acinetobacter baumannii* during a prolonged infection. *Microb. Genom.*, e000165.

Hnizda,A. *et al.* (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. *Leukemia*, **32**, 1393–1403.

Holt,K.E. *et al.* (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.*, **50**, 849–856.

Huang,Y. *et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Jafri,M. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.

Jankauskaite,J. *et al.* (2018) SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *bioRxiv*, 341735.

Jubb,H. *et al.* (2015) Flexibility and small pockets at protein-protein interfaces: new insights into druggability. *Prog. Biophys. Mol. Biol.*, **119**, 2–9.

Jubb,H.C. *et al.* (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.

Jubb,H.C. *et al.* (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. *Prog. Biophys Mol. Biol.*, **128**, 3–13.

Karmakar,M. *et al.* (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.*, **198**, 541–544.

Karmakar,M. *et al.* (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS One*, **14**, e0217169.

Kiyoshi,M. *et al.* (2014) Affinity improvement of a therapeutic antibody by structure-based computational design: generation of electrostatic interactions in the transition state stabilizes the antibody-antigen complex. *PLoS One*, **9**, e87099.

Nemethova,M. *et al.* (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. *Eur. J. Hum. Genet.*, **24**, 66–72.

Pandurangan,A.P. *et al.* (2017a) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.*, **45**, 303–311.

Pandurangan,A.P. *et al.* (2017b) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.*, **45**, W229–W235.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Phelan,J. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.*, **14**, 31.

Pires,D.E. and Ascher,D.B. (2016a) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic Acids Res.*, **44**, W557–561.

Pires,D.E. and Ascher,D.B. (2016b) mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.*, **44**, W469–W473.

Pires,D.E. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.

Pires,D.E. *et al*. (2014a) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*., **42**(Web Server issue), W314–319.

Pires,D.E. *et al*. (2014b) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.

Pires,D.E. *et al*. (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res*., **43**(Database issue), D387–391.

Pires,D.E. *et al*. (2016a) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep*., **6**, 29575.

Pires,D.E. *et al*. (2016b) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep*., **6**, 19848.

Portelli,S. *et al*. (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Sci. Rep*., **8**, 15356.

Ramdzan,Y.M. *et al*. (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. *Cell Rep*., **19**, 919–927.

Rodrigues,C.H. *et al*. (2018a) Kinact: a computational approach for predicting activating missense mutations in protein kinases. *Nucleic Acids Res*., **46**, W127–W132.

Rodrigues,C.H. *et al*. (2018b) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res*., **46**, W350–W355.

Rodrigues,C.H.M. *et al*. (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res*., **47**, W338–W344.

Roy,A. *et al*. (2017) In silico methods for design of biological therapeutics. *Methods*, **131**, 33–65.

Sefid,F. *et al*. (2019) In silico engineering towards enhancement of bap–VHH monoclonal antibody binding affinity. *Int. J. Pept. Res. Ther*., **25**, 273–287.

Silvino,A.C. *et al*. (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of plasmodium vivax relapses. *PLoS One*, **11**, e0160172.

Sinha,N. *et al*. (2002) Differences in electrostatic properties at antibody-antigen binding sites: implications for specificity and cross-reactivity. *Biophys. J*., **83**, 2946–2968.

Sirin,S. *et al*. (2016) AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Sci*., **25**, 393–409.

Soardi,F.C. *et al*. (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom. Med*., **2**, 7.

Stricher,F. *et al*. (2005) The FoldX web server: an online force field. *Nucleic Acids Res*., **33**(**Suppl_2**), W382–W388.

Tanaka,T. *et al*. (2014) Monoclonal antibodies in rheumatoid arthritis: comparative effectiveness of tocilizumab with tumor necrosis factor inhibitors. *Biologics*, **8**, 141–153.

Thiltgen,G. and Goldstein,R.A. (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One*, **7**, e46084.

Touw,W.G. *et al*. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*, **43**(Database issue), D364–368.

Traynelis,J. *et al*. (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res*., **27**, 1715–1729.

Trezza,A. *et al*. (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest. Ophthalmol. Vis. Sci*., **58**, 5320–5328.

Urquhart,L. (2018) Market watch: top drugs and companies by sales in 2017. *Nat. Rev. Drug Discov*., **17**, 232.

Usher,J.L. *et al*. (2015) Analysis of HGD Gene Mutations in Patients with alkaptonuria from the United Kingdom: identification of novel mutations. *JIMD Rep*., **24**, 3–11.

Van Regenmortel,M.H. (2014) Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. *J. Mol. Recognit*., **27**, 627–639.

Vedithi,S.C. *et al*. (2018) Structural implications of mutations conferring rifampin resistance in mycobacterium leprae. *Sci. Rep*., **8**, 5016.

Wager,S. *et al*. (2014) Confidence intervals for random forests: the Jackknife and the infinitesimal Jackknife. *J. Mach. Learn. Res*., **15**, 1625–1651.

Yugandhar,K. *et al*. (2017) PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, **33**, 2787–2788.

Appendix L

Computational Saturation Mutagenesis to predict structural consequences of systematic mutations on protein stability and rifampin interactions in the β subunit of RNA polymerase in *Mycobacterium leprae*

Computational and Structural Biotechnology Journal 18 (2020) 271–286

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of RNA polymerase in *Mycobacterium leprae*

Sundeep Chaitanya Vedithi [a,*], Carlos H.M. Rodrigues [b,c], Stephanie Portelli [b,c], Marcin J. Skwark [a], Madhusmita Das [d], David B. Ascher [a,b,c], Tom L. Blundell [a,*], Sony Malhotra [a,1]

[a] *Department of Biochemistry, University of Cambridge, Tennis Court Rd., CB2 1GA, UK*
[b] *Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Parkville, VIC 3052, Australia*
[c] *Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, VIC 3004, Australia*
[d] *Molecular Biology Laboratory, Schieffelin Institute of Heath-Research and Leprosy Center, Karigiri, Vellore, Tamil Nadu 632106, India*

ARTICLE INFO

Article history:
Received 18 July 2019
Received in revised form 3 January 2020
Accepted 7 January 2020
Available online 17 January 2020

Keywords:
Mutation Coolspots
Mycobacterium leprae
In-silico Saturation Mutagenesis
Thermodynamic stability
Rifampin
RNA Polymerase

ABSTRACT

Rifampin resistance in leprosy may remain undetected due to the lack of rapid and effective diagnostic tools. A quick and reliable method is essential to determine the impacts of emerging detrimental mutations in the drug targets. The functional consequences of missense mutations in the β-subunit of RNA polymerase (RNAP) in *Mycobacterium leprae* (*M. leprae*) contribute to phenotypic resistance to rifampin in leprosy. Here, we report *in-silico* saturation mutagenesis of all residues in the β-subunit of RNAP to all other 19 amino acid types (generating 21,394 mutations for 1126 residues) and predict their impacts on overall thermodynamic stability, on interactions at subunit interfaces, and on β-subunit-RNA and rifampin affinities (only for the rifampin binding site) using state-of-the-art structure, sequence and normal mode analysis-based methods. Mutations in the conserved residues that line the active-site cleft show largely destabilizing effects, resulting in increased relative solvent accessibility and a concomitant decrease in residue-depth (the extent to which a residue is buried in the protein structure space) of the mutant residues. The mutations at residue positions S437, G459, H451, P489, K884 and H1035 are identified as extremely detrimental as they induce highly destabilizing effects on the overall protein stability, and nucleic acid and rifampin affinities. Destabilizing effects were predicted for all the clinically/experimentally identified rifampin-resistant mutations in *M. leprae* indicating that this model can be used as a surveillance tool to monitor emerging detrimental mutations that destabilise RNAP-rifampin interactions and confer rifampin resistance in leprosy.

Author summary: The emergence of primary and secondary drug resistance to rifampin in leprosy is a growing concern and poses a threat to the leprosy control and elimination measures globally. In the absence of an effective *in-vitro* system to detect and monitor phenotypic resistance to rifampin in leprosy, diagnosis mainly relies on the presence of mutations in drug resistance determining regions of the *rpoB* gene that encodes the β-subunit of RNAP in *M. leprae.* Few labs in the world perform mouse footpad propagation of *M. leprae* in the presence of drugs (rifampin) to determine growth patterns and confirm resistance, however the duration of these methods lasts from 8 to 12 months making them impractical for diagnosis. Understanding molecular mechanisms of drug resistance is vital to associating mutations to clinically detected drug resistance in leprosy. Here we propose an *in-silico* saturation mutagenesis approach to comprehensively elucidate the structural implications of any mutations that exist or that can arise in the β-subunit of RNAP in *M. leprae.* Most of the predicted mutations may not occur in *M. leprae* due to fitness costs but the information thus generated by this approach help decipher the impacts of mutations across the structure and conversely enable identification of stable regions in the protein that are least impacted by mutations (mutation coolspots) which can be a potential choice for small molecule binding and structure guided drug discovery.

* Corresponding authors.
E-mail addresses: scv26@cam.ac.uk (S.C. Vedithi), tlb20@cam.ac.uk (T.L. Blundell).
[1] Present address: Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, London WC1E 7HX, UK.

https://doi.org/10.1016/j.csbj.2020.01.002

1. Introduction

Nonsynonymous mutations in genes that encode drug targets in mycobacteria can induce structural and consequent functional changes leading to antimicrobial resistance, the burden of which is rapidly increasing and is a global health concern. Diagnosis of ~600,000 new cases of rifampin-resistant tuberculosis in 2018 suggest that it poses a risk for the concomitant increase in undiagnosed rifampin-resistant leprosy, worldwide [1]. *Mycobacterium leprae (M. leprae),* the causative bacilli for leprosy, is phylogenetically closest to *Mycobacterium tuberculosis* [2] and developed resistance to rifampin before the introduction of World Health Organization (WHO) recommended multi-drug therapy (MDT) in the year 1984. Despite the long duration of chemotherapy with MDT (six months in paucibacillary to 12 months in multibacillary disease), rifampin-resistant case numbers are less and represent only 3-5% of total clinically diagnosed relapsed leprosy cases as reported by WHO in 2017 [3]. One of the possible reasons for the low numbers of drug-resistant leprosy cases globally is the lack of quick, effective and reliable *in-vitro* diagnostic test for confirming phenotypic resistance. Current methods rely on identifying mutations in the rifampin resistance determining region (RRDR) of the *rpoB* gene through gene sequencing and/or by testing growth patterns of *M. leprae* in response to individual drugs in the MDT in an *in-vivo* mouse footpad model; however, the later technique is both time and labour intensive.

While mutations within the β-subunit of RNAP contribute to clinical resistance to rifampin, the associated structural changes can complicate the transcription process in bacteria by modulating various physiological processes [4], the knowledge of which is essential for novel drug discovery or alternative therapies to treat rifampin resistant strains of *M. leprae*. In the absence of an artificial culture system to propagate and study mechanisms of resistance, it is exceptionally challenging to define an experimental phenotype for rifampin resistance in leprosy. *M. smegmatis* as a surrogate host with cloned *M. leprae rpoB* gene has proved a dependable model to study phenotypic effects; however, this technique is limited to biosafety level-2 laboratories that have facilities for gene cloning and sequencing, and cannot be translated to a regular diagnostic setting in leprosy endemic countries [5]. A plausible association between mutations in drug targets and phenotypic resistance outcomes could be established if minimum inhibitory concentrations (MICs) of the drugs are known for the mutant strains. While MICs can be estimated in cultivable species like *M. tuberculosis* and *M. smegmatis*, obtaining growth information from *in vivo* propagation for a slow growing and obligate pathogen like *M. leprae* is challenging and needs time and resources. Alternatively, *in-silico* methods that predict structural implications of mutations can be useful in understanding mechanisms of resistance and help prioritise mutations that require experimental validation in leprosy, owing to the absence of a tool for quantitative estimation of phenotypic resistance [6].

Mutations contribute to disruption of protein–ligand and protein-nucleic acid interactions resulting in drug resistance in mycobacterial diseases [7,8]. Changes in affinity between the drug target protein and the ligand can result from both orthosteric and allosteric mechanisms leading to various resistance phenotypes [4]. The β-subunit of RNAP in *M. leprae* is encoded by the *rpoB* gene (ML1891) whose product is 1178 amino acids in length. The RRDR is located between the residue positions 410 and 480. Approximately 40 mutations have been reported in the *rpoB* gene of *M. leprae* that induce clinical resistance to rifampin in leprosy [9–11]; however, in tuberculosis, nearly 100 mutations have been reported in the same gene that shares 96% gene sequence identity with that of *M. leprae* [12]. As the burden of rifampin resistance is very high in *M. tuberculosis* with known and new mutations being reported from different studies [13–17], it is important to monitor the emergence of new rifampin-resistant mutations in *M. leprae*. A comprehensive understanding of the effect of any mutation on the structure of RNAP is vital in the context of monitoring emerging rifampin resistance and its implications on controlling global leprosy incidence.

In order to decipher the effect of systematic mutations on the stability of the protein structure, protein sub-unit interfaces, nucleic acid and ligand interactions, we performed *in-silico* saturation mutagenesis (mutating every residue to all the other 19 residues) and predicted the change in stability of the β-subunit and affinity between β-subunit andrest of the subunits in the complex, β-subunit-rifampin and β-subunit-RNA interactions. Additionally, we also assessed the impacts of mutations on the secondary structures of the polypeptide chains, relative sidechain solvent accessibility, residue-depth and residue-occluded packing density [18]. Residue-level evolutionary conservation scores were determined and compared with the predicted destabilizing effects. Extremely detrimental mutations (that destabilize β-subunit of RNAP and affinity between β-subunit -rest of the subunits in the complex, β-subunit -rifampin and β-subunit-RNA interactions) were selected and analysed for changes in their interatomic interactions that might explain the reasons for the predicted destabilizing effects. To explore further, the vibrational entropy and enthalpy changes of the protein in flexible conformations, we employed an empirical force field-based method – FoldX [19], a course-grained normal mode analysis (NMA) based elastic network contact model – ENCoM [20] and a consensus predictor that integrates normal mode approaches with graph-based distance matrix in the mutating residue environment– DynaMut [21]. Finally, fragment hotspots [22] were mapped on the structures to provide information on potential druggable sites whose stability is predicted to be least likely affected by mutations (no mutations in these regions were identified in leprosy). We termed these sites as "Mutation coolspots" which can be explored for novel/alternative small molecule binding and structure-guided drug discovery to treat rifampin-resistant leprosy.

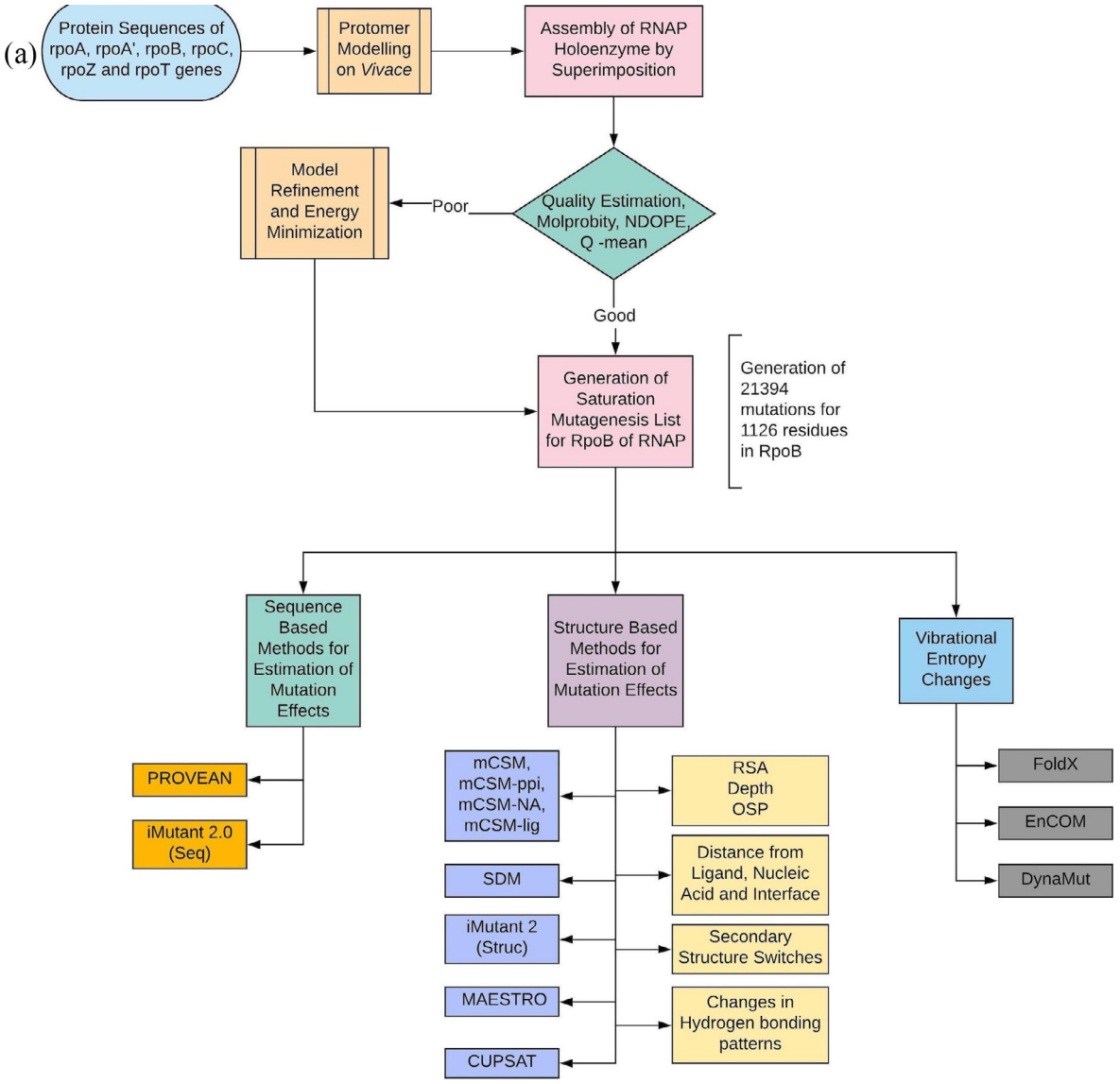
2. Materials & Methods

2.1. Design:

The key stages in the methodology involve comparative protein 3D modelling using known crystal structures of homologues as templates, quality assessment of the built models, generating mutation lists from the model and sequential submission of the lists and the model to stability change prediction servers for sequence, structure and vibrational entropic terms (Fig. 1A).

2.2. Comparative modelling, quality assessment and model refinement:

A model for RNAP holoenzyme of *M. leprae* was built using Modeller 9.21 [23] with templates from *M. tuberculosis* (PDB Id:5UH5 (96% identity, 3.74 Å resolution) containing RNAP, nucleic acid scaffold with DNA and three nucleotides of RNA complementary to the template DNA strand, and PDB Id: 5UHC (96% identity, 3.79 Å resolution) containing all the elements similar to 5UH5 and rifampin) as described earlier by us [4]. The quality of the generated model was assessed using Molprobity [24] and atomic clashes were removed by minimizing the energy of the model by 100 steps using Steepest Decent (step size = 0.02 Å) and by 10 steps (step size = 0.02 Å) using conjugate gradient methods. Energy minimizations were performed using UCSF Chimera [25]. The mutant models were generated using a script from Modeller 9.21



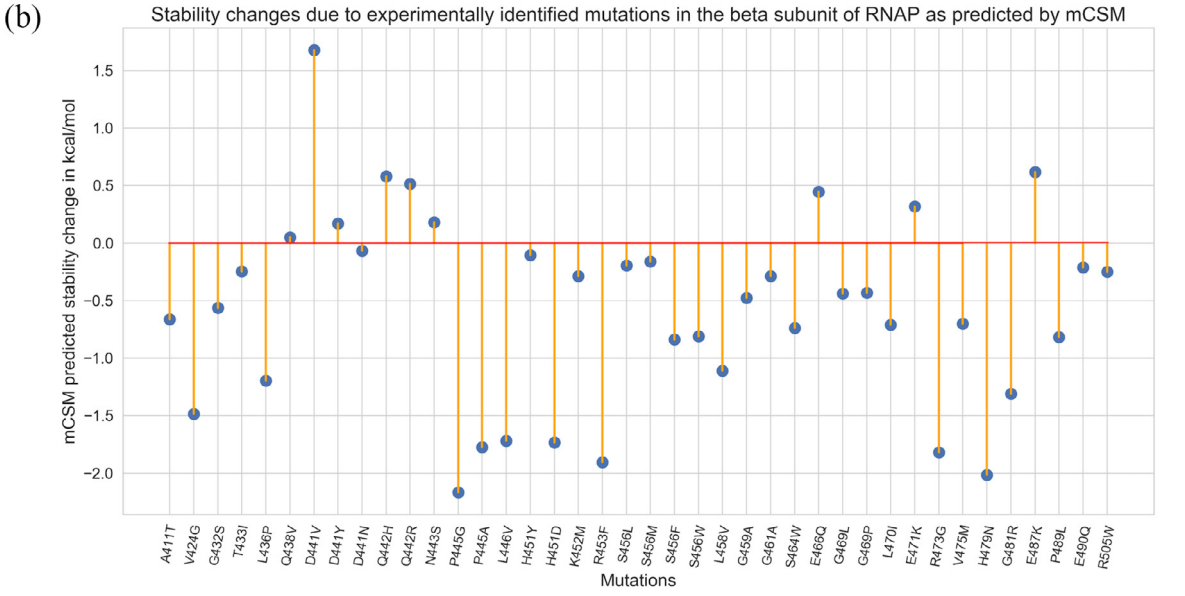


Fig. 1. [A] Methodology and study design. [B] A lollipop plot with stability predictions for mutations reported in the literature and are known to confer rifampin resistance in Leprosy.

(mutate_model.py) and sidechains of the mutants were optimized using ANDANTE [26], a program that uses χ angle conservation criteria to optimize the sidechain rotamers. Multiple models were generated initially to test the variation in the modelling process. Structural similarity among the models was tested using root mean square deviation (RMSD) and TM-Align scores [27]. (Supplementary Figs. 4–6, and Supplementary Table 3).

2.3. Saturated Mutagenesis:

A systematic list of 21,394 mutations was generated for residues starting from P28 and ending at E1153 positions in the β-subunit (the modelled region). This list was programmatically submitted to a set of servers as stated in Table 1 below:

2.4. Residue conservation:

Conservation scores for each residue in the wild-type model was estimated using CONSURF – a server that uses evolutionary patterns of amino acids/nucleic acids from the multiple sequence alignment and develops a probabilistic framework to calculate evolutionary rates for each residue in the sequence.

2.5. Effects of mutations on protein stability and interactions:

The effect of mutations on thermodynamic stability of the β-subunit of RNAP was analyzed using mCSM, SDM and FoldX4. For SDM, mutant models were generated using ANDANTE. The effect of mutations on RNA affinity is assessed using mCSM-NA2 on mutant models with nucleic acid scaffold. The holoenzyme complex of RNAP consists of five subunits and the effects of mutations on the protein–protein interfaces (between β and all the other sub-units in RNAP complex) were assessed using mCSM-ppi. Rifampin binds to the β-subunit of RNAP and we analyzed the effects of mutations on the protein–ligand affinity using mCSM-lig. Only residues that are within 10 Å of interatomic distance to rifampin were analyzed by mCSM-lig.

The stability changes were further compared with predictions from other sequence- (PROVEAN, I-Mutant 2.0 (Sequence) and structure-based (MAESTRO, CUPSAT, I-Mutant 2.0 (Structure)) computational tools in order to estimate the reliability of the predictions.

2.6. Changes in vibrational entropy and normal mode analysis:

In order to determine the effects of the mutations in flexible conformations of the protein, we used FoldX4, an empirical force field approach that calculates free energy changes between native and mutant forms of the protein, and an elastic network contact model (ENCoM), which is a coarse grain NMA method that considers the nature of the amino-acids and aids in calculating vibrational entropy changes upon mutations. We also used DynaMut, a consensus predictor of protein stability based on the vibrational entropy changes predicted by ENCoM and the stability changes predicted by graph-based signatures that are used in mCSM program.

2.7. Conformational changes:

Conformational changes and their impacts on biophysical properties of the proteins were estimated using SDM. The interatomic distances between each residue and the interface with other subunits in the RNAP holoenzyme, rifampin and nucleic acids in the structure were measured and included in the analysis. Secondary

Table 1
List of servers used in the computational analysis:

Si No:	Name of web server	Function	Reference	Submission parameters
1	mCSM	Predict protein stability changes due to mutations.	[28]	Model PDB file, mutation and chain id.
2	SDM	Predict protein stability changes due to mutations.	[18]	Model PDB file, mutation and chain id.
3	mCSM-PPI	Predict stability of protein–protein interfaces due to mutations.	[28]	Model PDB file, mutation and chain id.
4	mCSM-NA2	Predict stability of protein-nucleic acid interactions due to mutations	[29]	Model PDB file, mutation, chain id and nucleic acid type.
5	mCSM-lig	Stability of protein–ligand interactions due to mutations	[30]	Model PDB file, mutation, chain id, three letter code of the ligand and ligand affinity in wild type structure in nM concentration.
6	FoldX4	Predict protein stability changes due to mutations.	[19]	Model PDB file, list of mutations and chain ids.
7	MAESTRO	Predict protein stability changes due to mutations.	[31]	Model PDB file, list of mutations and chain ids.
8	CUPSAT	Predict protein stability changes due to mutations.	[32]	Model PDB file, list of mutations and chain ids.
9	Imutant 2.0-Struc	Predict protein stability changes due to mutations.	[33]	Model PDB file, list of mutations and chain ids.
10	Imutant 2.0 -Seq	Predict protein stability changes due to mutations using sequence information.	[33]	RNAP sequence file in fasta format, list of mutations and chain ids.
11	PROVEAN	Predict protein stability changes due to mutations using sequence information.	[34]	RNAP sequence file in fasta format, list of mutations and chain ids.
12	CONSURF	To calculate evolutionary conservation score of each residue in the protein.	[35]	Model PDB file
13	ENCoM	Conformational Changes in protein due to mutations.	[20]	Model PDB file, list of mutations and chain ids.
14	DynaMut	Conformational Changes in protein due to mutations.	[21]	Model PDB file, list of mutations and chain ids.
15	Arpeggio	Map interatomic interactions between wildtype and mutant amino acids and the residue environment.	[36]	Model PDB file and the residue selection in standard format.
16	Intermezzo	Map interatomic interactions between wildtype and mutant amino acids and the residue environment.	Bernardo Ochoa Montano & Blundell TL unpublished	Model PDB file and the residue selection in standard format.
17	ANDANTE	Works along with Modeller to generate mutant models from wildtype model files.	[26]	Model PDB file, mutation and chain id
18	Fragment Hotspot Maps	Maps regions on the surface of the protein that has high propensity for small molecule binding.	[22]	Model PDB file.

structure switches in mutants, changes in relative solvent accessibility, depth of the residue in Å and residue-occluded packing densities were determined for all the mutations.

2.8. Interatomic interactions:

After predicting protein stability changes and changes in RNAP-rifampin affinities, mutations at two positions vide H451 & P489 that highly destabilize rifampin binding and are experimentally identified in the rifampin resistant leprosy patients [9,10] (present in the set of 40 experimentally identified mutations – Supplementary Table 2), were analyzed for the changes in interatomic interactions of the mutating residues using Arpeggio, a program that maps the types of interatomic interactions of wildtype and mutant residues with the residue environment based on atom type, interatomic distance and angle constraints. Additionally, four mutations at positions S437, G459, K884 & H1035 which are computationally predicted to highly destabilize RNAP-rifampin interactions were chosen and subjected to similar analysis. Intermezzo program (Bernardo Ochoa Montano & Blundell TL unpublished) was also used for interactive analysis of bonding patterns on Pymol sessions.

2.9. Fragment hotspot maps:

Fragment hotspot maps aid in locating specific sites on the surface of the protein that are topologically, chemically and entropically favorable for small molecule (fragment) binding. The atomic hotspots on the drug target are explored computationally using donor, acceptor and hydrophobic fragment probes, and introducing a depth criterion to assist in estimating the small molecule binding propensity. For ligand-binding proteins, the fragment hotspot maps aid in understanding the pharmacophore characteristics of the interacting regions. We mapped the hotspots on the β-subunit of RNAP and colored the surface with regions that are least impacted by any mutations (mutation coolspots).

3. Results

In total, 21,394 mutations were generated from 1126 residues in the β-subunit of RNAP (Supplementary Table 1). The list of experimentally identified mutations and their effects are separately shown in Supplementary Table 2.

3.1. Multivariate analysis of free energy change predictions by various computational tools for saturated mutations:

Along with the in-house developed mCSM and SDM tools for prediction of protein stability changes upon saturated mutagenesis of the β-subunit of RNAP, a comparative analysis was performed with other sequence (PROVEAN, I-mutant 2.0 – Sequence), structure- (CUPSAT, I-mutant 2.0-structure, MAESTRO) and NMA-based tools (FOLDX, ENCOM, DynaMut). Average stability changes caused by all possible mutations at each residue position in the β-subunit of RNAP, as predicted by mCSM and SDM, were compared with other structure-based predictors (Supplementary Fig. 1) (rifampin-interacting residues are highlighted). Correlation of overall stability predictions performed by mCSM with each of the other tools indicated an "r" value of 0.55 with SDM, 0.61 with MAESTRO, 0.72 with Imutant 2.0 (Structure) and 0.43 with CUPSAT. Correlations between mCSM, SDM and other sequence and NMA based tools are shown in Supplementary Figs. 2 and 3. The rationale for performing these correlations is to understand how mCSM and SDM being structure-based predictors of stability changes, relate to sequence-based methods and vibrational entropy changes in normal mode perturbations.

3.2. Experimentally/Clinically identified mutations:

We performed a systematic literature review to list all the mutations reported in the β-subunit of RNAP in *M. leprae*. We noted 40 mutations at 32 unique residue positions. The reference articles are listed in Supplementary Table 2. As depicted in Fig. 1B, 77.5% [19] of the experimentally/clinically identified mutations destabilize the β-subunit. Except for A411T and V424G mutations, all the other residues are present in close proximity to rifampin binding sites (Fig. 2A) and destabilize rifampin interactions (as predicted by mCSM-lig).

3.3. Residue conservation and protein stability:

The stability changes, predicted after saturation mutagenesis of each residue in the β-subunit, were compared with residue conservation scores. CONSURF scores of less than zero are attributed to conserved residues and scores of zero and above to variable residues (score 3 being highly variable). The average change in protein stability that was predicted by mCSM for mutations at each residue position ranged from 0.823 to −3.033 kcal/mol and that of SDM varied from 2.167 to −4.36 kcal/mol. Residues that line the active center cleft and interact with rifampin and the nucleic acid scaffold are highly conserved, while surface exposed residues have variable conservation scores (Fig. 2B). Rifampin-interacting residues between positions ~400–500 are highly conserved and 87.3% of the saturated mutations in this region destabilize the protein (Supplementary Table 1). The maximum destabilizing effect of mutations at each of these residues varied between −0.311 to −4.311 kcal/mol (mCSM). The maximum destabilizing mutation is defined as a mutation that induces a maximum decrease in Gibbs free energy (stability change) of the β-subunit of RNAP, RNAP-rifampin and RNAP-subunit interactions among all the 19 possible mutations at each residue position (when predicted by mCSM, SDM, mCSM-lig and mCSM-ppi software). The maximum destabilizing effect predicted by mCSM for all possible mutations at each residue was mapped on the structure to identify regions that are largely impacted by mutations (Fig. 2C). Conversely, the residues whose stability is least impacted by all possible mutations are colored in blue to identify "mutation coolspots" that are potentially areas of choice for targeting with small molecules in drug discovery (Fig. 2D).

As part of the RNAP holoenzyme complex, the β-subunit interacts with other subunits and has large interfacial regions. The impact of mutations on the stability of these interfaces was measured using mCSM-PPI. It was noted that the maximum destabilizing effect by any mutation at a particular residue in the interface between β and β′ subunits has an affinity change that ranged from −0.021 to −5.108 kcal/mol (−5.108 kcal/mol was noted for mutation W1074R which is not reported experimentally in rifampin resistant leprosy cases). The interfacial region and the stability changes are mapped on the structure (Fig. 3A and B).

3.4. Relative sidechain solvent accessibility (RSA), residue-depth, residue-occluded packing density and protein stability:

The difference in relative solvent accessibility between wild type and the mutant residues for all the mutations were calculated using SDM. While analyzing the maximum destabilizing mutations among all the possible mutations at each residue position, it was noted that maximum destabilizing mutants at 751 residue positions (66.79%) showed increase in RSA. The maximum destabilizing mutants at rest of the 375 positions indicated a decrease in RSA.

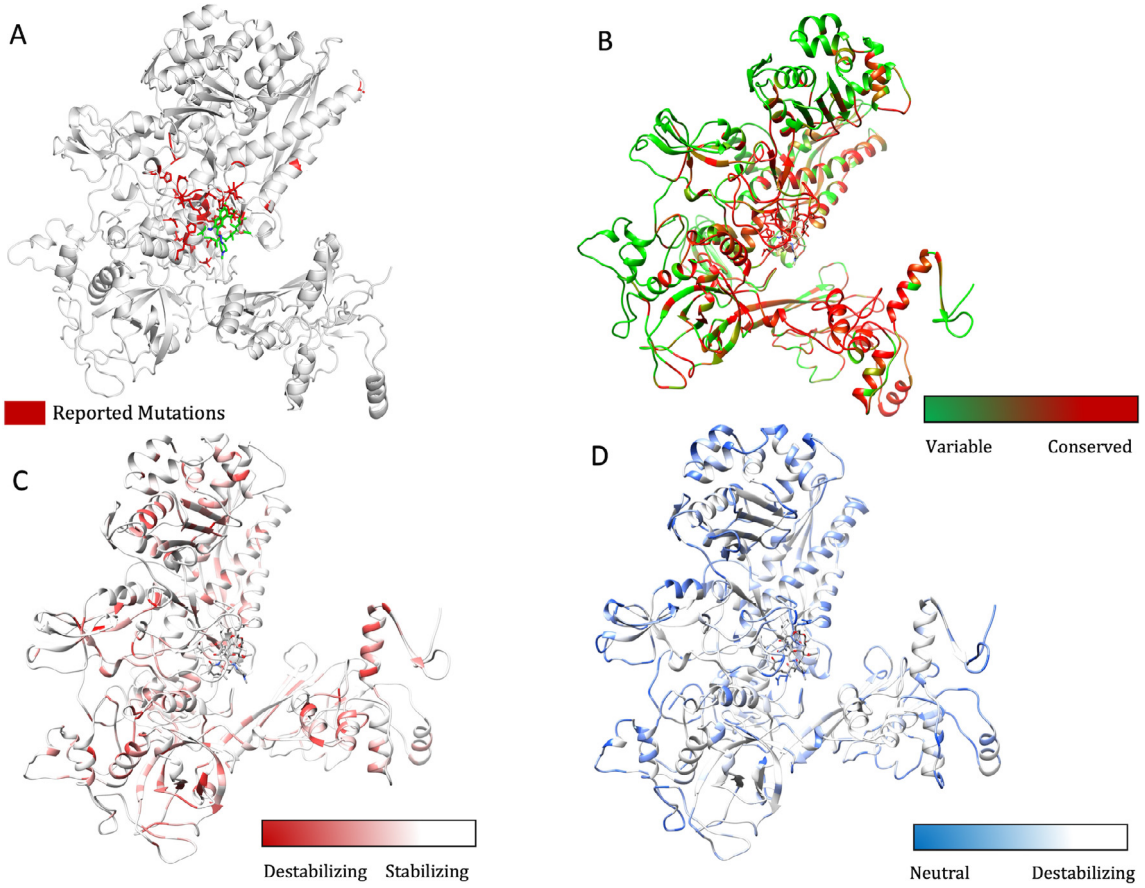


Fig. 2. [A] The β-subunit of RNAP with residues where mutations were reported experimentally from patient samples in various studies (Supplementary Table 2) (highlighted in red). [B] Each residue in the β-subunit of RNAP that is colored by the conservations scores determined by CONSURF. The residues in green are variable (conservations scores greater than 1) and are usually surface exposed. The residues in red are conserved with conservation scores less than 1 and usually form the core of the protein. The rifampin binding site is highly conserved in *M. leprae*. [C] The maximum destabilizing effect (predicted by mCSM) on the protein stability for any mutation at each residue position, is mapped on the structure. Red are the regions that are largely destabilized by mutations while the white regions are relatively stable with mutations. [D] The converse of B where the regions, whose stability is least impacted by mutations, are coloured in blue and we called them "Mutation CoolSpots". (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

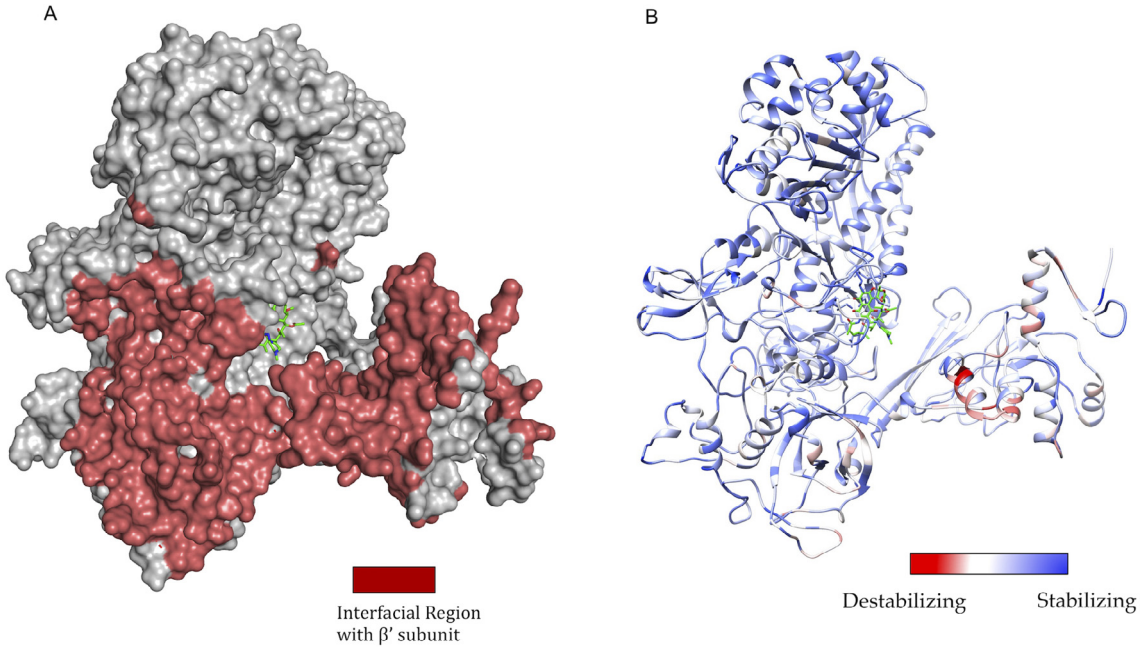


Fig. 3. [A] The interfacial region of the β-subunit of RNAP highlighted in Maroon. [B]. The maximum destabilizing effect a mutation can induce on the interface stability, is predicted by mCSM-PPI and mapped on the structure. Red indicates regions that are highly destabilized by mutations (-5.108 Kcal/mol) while the blue indicates stable regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Among the 751 mutants with increase in RSA, 551 were hydrophobic and 121 substitutions within 551 were from polar/charged (wildtype) to hydrophobic residues (mutants). As mutant hydrophobic residues with increased solvent accessibility often destabilize the protein [38], the destabilizing effects of these mutations ranged from −1.021 to −4.311 kcal/mol. Additionally, these substitutions resulted in a decrease in residue-depth [18] (ranging from 0.01 Å to 1.83 Å), which is concomitant with the increase in solvent accessibility. These changes in RSA and depth at the rifampin-binding site are depicted in Fig. 4A and B.

From the maximum destabilizing mutations at all the 1126 positions, mutations at 586 (52.04%) residue positions resulted in increase in residue-depth that ranged from 0.01 to 2.46 Å. Mutants were generated using ANDANTE which places the side chains without any steric clashes and the mutant models were subjected to energy minimization. Hence the change in residue-depth is attributed to the buriedness of the residue and not just the natural change from a larger to a smaller amino acid. The decrease in residue-depth in the remaining 540 (47.95%) residues ranged from 0.1 to 3.02 Å. Similarly, the residue-occluded packing density [18] increased at 539 residue positions (47.86%). These changes in RSA and residue-depth are mapped as attributes on to the structure of the β-subunit of RNAP and it was noted that most of the residues that line the active center cleft have increase in RSA upon mutations. Decrease in residue-depth was noted in residues at the rifampin-binding pocket and at the subunit interfaces (Fig. 5A and B).

3.5. Substitutions to aspartate predominate mutations that destabilize the β-subunit-RNA affinity in RNAP:

The effects of mutations on β-subunit-RNA affinity was estimated using mCSM-NA2. Substitutions to aspartate residues were most common among mutations that highly destabilize β-subunit-RNA interactions in RNAP. The mutant aspartate residues form π-π interactions with the nucleotides in RNA either by stacking or by nucleotide-edge T-shaped and amino-edge T-shaped interactions. Aspartate being an acyclic π-containing amino acid, readily forms nucleotide (edge) amino (edge) or nucleotide (face) and amino-acid (edge) interactions [39]. This ability of acyclic amino acids like arginine, glutamic acid and aspartic acid to form a variety of charged-π interactions with nucleotides in mutants may impact the orientation of RNA molecules in the active center cleft of RNAP leading to loss or gain in function. Approximately, 93% of the highly destabilizing mutations at RNA-interacting residues are substitutions to aspartate. Mutations to glutamate were also noted in 6.83% and additionally one each of methionine, proline and threonine mutations indicated highly destabilizing effects.

3.6. Substitutions to arginine predominate mutations that destabilize β-subunit-rifampin affinity:

Systematic mutations in the set of 70 residues that lie 10 Å from the rifampin binding site reveal that mutations that largely destabilize RNAP-rifampin affinities are primarily arginine and glutamate substitutions (mCSM-lig). In the binding site, R173, R454, R465 and R613 form hydrogen bonds and a network of other interactions with rifampin that stabilize the molecule in the binding site [4]. Introduction of additional arginine residues by mutations may influence the stability and orientation of rifampin in the binding site. The positively charged guanidinium ion of arginine forms cation-π interactions with aromatic amino acids as noted in earlier studies [40,41]. In the predicted mutations S437R and G456R, arginine forms an intricate network of π interactions with surrounding aromatic amino acids changing the shape of the binding pocket and leading to a loss in rifampin interactions (rifampin retains only two polar contacts with Q438 and F439 whereas wildtype has five hydrogen bonds). The effects of mutations on RNA and rifampin affinity as predicted by mCSM-NA2 and mCSM-lig were mapped on to the structure (Fig. 6A and B).

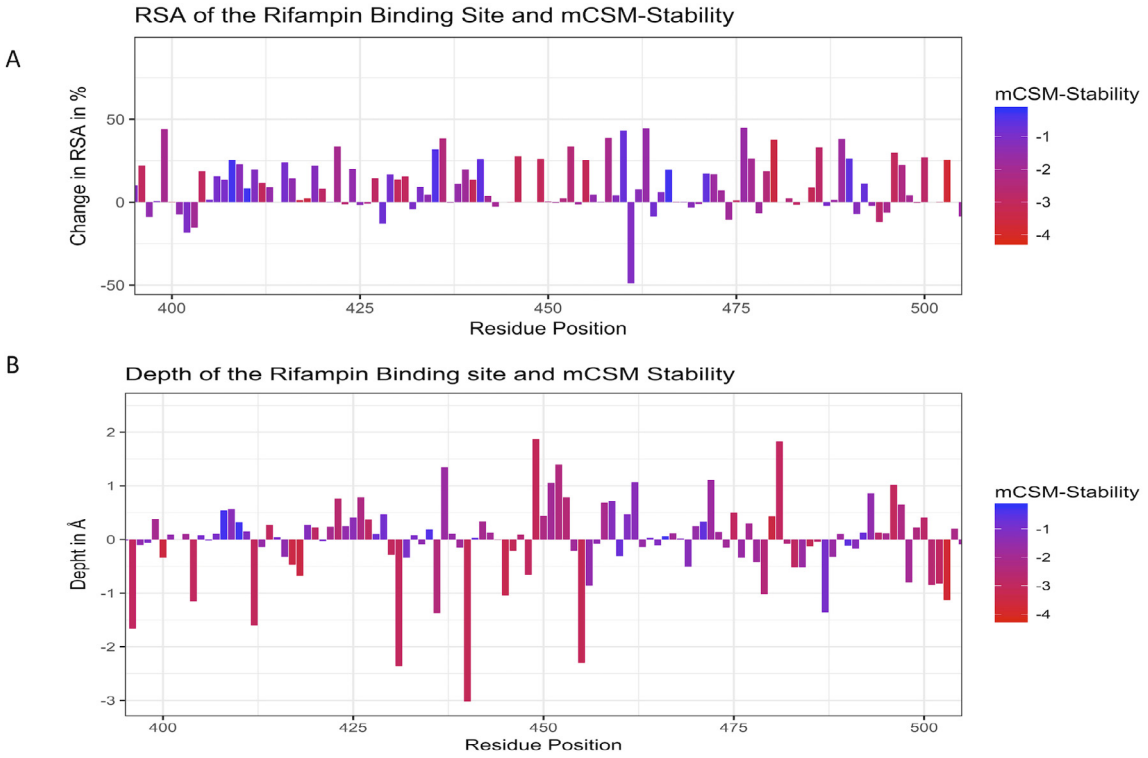


Fig. 4. [A] Change in relative solvent accessibility for maximum destabilizing mutants in the rifampin binding pocket (mCSM). [B]. Change in depth of the highly destabilizing mutant residue in the rifampin binding pocket (mCSM).

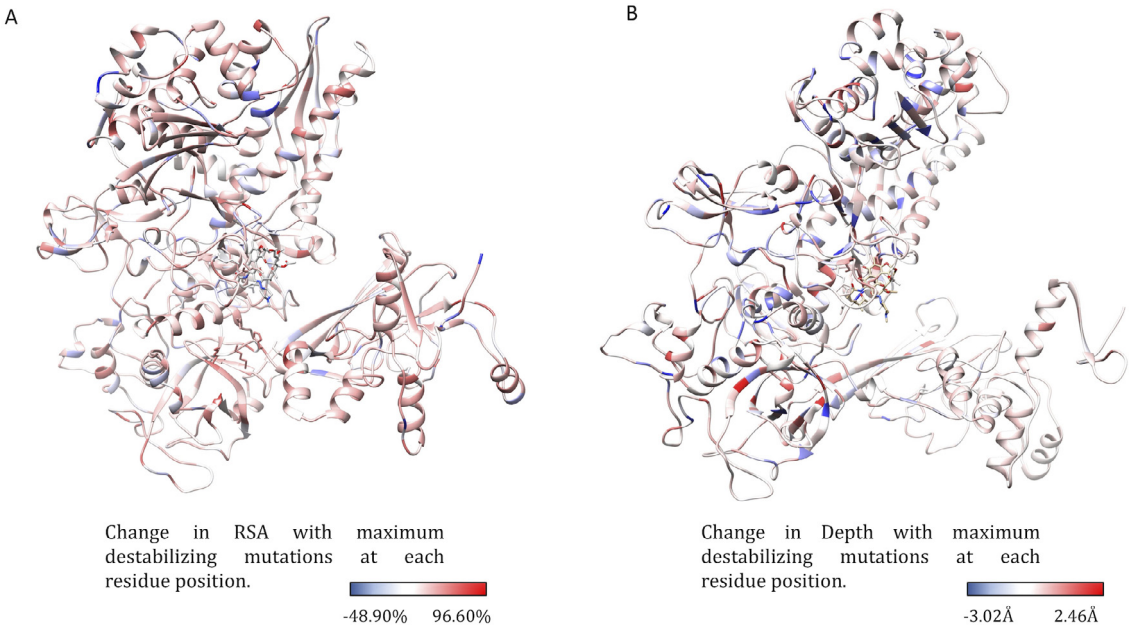


Fig. 5. [A] The change in relative side chain solvent accessibility with mutations was mapped on to the structure. Blue indicates a decrease in RSA while red indicates an increase. [B] The changes in depth with highly destabilizing mutations at each residue position was mapped on the structure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

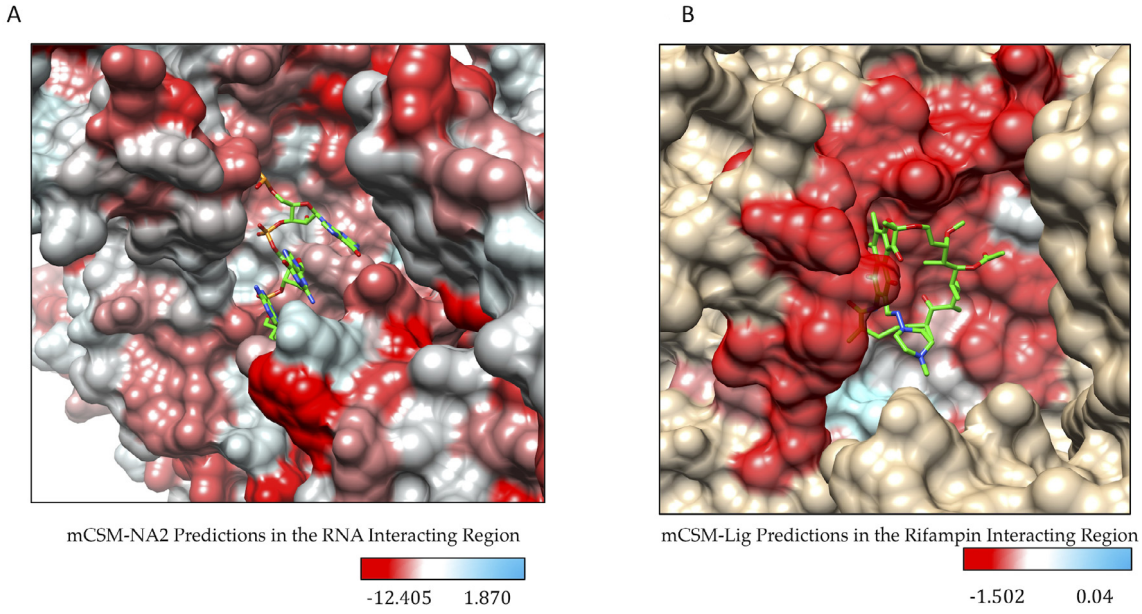


Fig. 6. [A] Stability changes in β-subunit -RNA and β-subunit- rifampin [B] interactions due to mutations in the binding sites as predicted by mCSM-NA2 and mCSM-lig. The maximum destabilizing effect a mutation can cause at each residue position in the binding site is depicted on the structure.

To determine if mCSM-lig predicted RNAP-rifampin binding affinities can provide information on the degree of resistance associated with each mutation in the rifampin binding site, we attempted to correlate MIC values of *M. tb rpoB* mutants with mCSM-lig predictions for RNAP-rifampin affinity in the structure of *M. tb* (PDB Id: 5UHC). A total of 40 mutations were selected from two studies [12,42] and mCSM-lig predictions were correlated with MIC values. It was noted that mCSM-lig predictions were independent of the MIC values which was also observed in an earlier study [43]. A table with MIC values and corresponding mCSM-lig predictions was included in Supplementary Material S1 (Table SM1). Additionally, a table with saturated mutations for all residues within 10 Å of the rifampin and their mCSM-lig predictions were presented in Supplementary Material S1 (Table SM2).

3.7. Detrimental mutations:

Among all the experimentally identified and computational predicted mutations, we selected those that highly destabilize (maximum decrease in log affinity fold change among all 19 mutations at each residue position) RNAP-rifampin interactions. Six residues were chosen based on the following characteristics and the structural effects of systematic mutations at each residue position were analyzed (Table 2) as below:

Table 2
Detrimental mutations and their corresponding stability changes that influence holoenzyme assembly, rifampin and RNA interactions.

Method	Wild-type residue	Residue position	Average stability effect	Maximum stabilizing effect	Mutant residue	Maximum destabilizing effect	Mutant residue
mCSM-Stability (ΔΔG in kcal/mol)	S	437	−0.795	−0.072	L	−1.701	H
	H	451	−1.214	−0.104	Y	−1.898	S
	G	459	−0.713	−0.381	V	−1.201	W
	P	489	−1.135	−0.507	R	−1.771	G
	K	884	−1.227	−0.190	L	−2.298	S
	H	1035	−0.419	0.600	Y	−1.421	G
mCSM-ppi (ΔΔG in kcal/mol)	S	437	−0.254	0.395	H	−0.820	R
	H	451	−0.652	−0.050	S	−1.451	M
	G	459	−0.397	0.237	H	−1.042	R
	P	489	−0.738	−0.138	W	−1.372	R
	K	884	−0.105	0.160	D	−0.685	R
	H	1035	−0.754	0.115	W	−1.726	R
mCSM-NA2 (ΔΔG in kcal/mol)	S	437	−1.538	4.922	W	−3.857	D
	H	451	−1.300	5.147	W	−3.632	D
	G	459	2.289	8.556	W	−0.221	D
	P	489	1.926	8.195	W	−0.582	D
	K	884	0.221	6.647	W	−2.130	D
	H	1035	0.847	7.295	W	−1.484	D
mCSM-lig (log-affinity change)	S	437	−0.646	−0.484	L	−1.062	R
	H	451	−0.510	−0.076	W	−0.777	E
	G	459	−0.981	−0.715	A	−1.236	R
	P	489	−0.598	−0.254	L	−0.917	R
	K	884	−0.156	−0.368	D	−0.925	R
	H	1035	−0.121	0.097	V	−0.501	E
SDM (ΔΔG in kcal/mol)	S	437	0.087	2.320	V	−1.900	P
	H	451	−0.756	1.290	L	−2.800	G
	G	459	−2.842	−1.780	V	−3.800	P
	P	489	−0.432	1.440	Y	−1.070	E
	K	884	0.108	1.270	V	−1.820	P
	H	1035	−0.200	0.590	V	−1.410	P
MAESTRO (ΔΔG in kcal/mol)	S	437	−0.21	−0.14	K	0.24	F
	H	451	−0.12	−0.05	G	0.22	R
	G	459	−0.23	−0.17	S	0.33	W
	P	489	−0.26	−0.22	H	0.31	M
	K	884	−0.20	−0.14	G	0.25	M
	H	1035	−0.27	−0.25	P	0.31	Y
CUPSAT (ΔΔG in kcal/mol)	S	437	2.70	7.98	I	−1.12	G
	H	451	2.01	6.92	W	−3.25	K
	G	459	−2.51	5.00	K	−5.53	C
	P	489	−2.76	−0.84	A	−5.47	M
	K	884	−2.99	3.42	I	−8.03	H
	H	1035	−1.07	2.15	C	−3.23	Y
Imutant 2.0 Structure (Sign of prediction)	S	437	4.05	9.00	A	1.00	F
	H	451	6.00	8.00	G	3.00	L
	G	459	6.63	9.00	N	3.00	I
	P	489	7.11	9.00	G	3.00	L
	K	884	6.42	9.00	G	2.00	M
	H	1035	4.63	8.00	G	2.00	L
PROVEAN (ΔΔG in kcal/mol)	S	437	−4.79	−3.00	A	−7.00	W
	H	451	−8.66	−5.73	Y	−10.37	C
	G	459	−8.10	−6.00	A	−10.00	L
	P	489	−9.04	−7.99	A	−10.99	F
	K	884	−5.97	−2.91	R	−7.75	C
	H	1035	−8.98	−5.79	Y	−10.61	C
Imutant 2.0 Sequence (Sign of prediction)	S	437	4.47	7.00	F	0.00	H
	H	451	3.21	7.00	P	0.00	F
	G	459	3.53	7.00	H	0.00	A
	P	489	6.89	9.00	G	5.00	L
	K	884	3.53	8.00	V	0.00	G
	H	1035	2.95	6.00	G	0.00	V
FOldX4 (ΔΔG in kcal/mol)	S	437	2.79	−1.44	I	12.39	R
	H	451	1.78	−0.74	L	4.39	W
	G	459	9.14	3.96	A	20.76	H
	P	489	3.04	2.11	N	4.79	R
	K	884	1.06	−2.12	Y	9.77	L
	H	1035	0.77	−1.47	P	5.69	Y

(*continued on next page*)

Table 2 (*continued*)

Method	Wild-type residue	Residue position	Average stability effect	Maximum stabilizing effect	Mutant residue	Maximum destabilizing effect	Mutant residue
ENCoM (ΔΔSvib in kcal/mol/K)	S	437	−0.44	0.48	G	−1.50	W
	H	451	0.34	0.97	G	−0.46	W
	G	459	−0.91	−0.29	A	−1.55	W
	P	489	−0.16	0.14	G	−0.82	F
	K	884	0.18	0.96	G	−0.60	W
	H	1035	0.19	0.73	G	−0.26	W
DynaMut (ΔΔG in kcal/mol)	S	437	2.87	6.99	L	−2.08	G
	H	451	−0.74	2.17	Y	−3.43	T
	G	459	1.93	3.29	N	−0.25	S
	P	489	0.94	3.26	F	−0.72	S
	K	884	0.14	3.69	W	−1.87	E
	H	1035	0.21	2.38	W	−2.29	G

- Mutations that highly destabilize rifampin binding (at wildtype S437 & G459 positions) as predicted by mCSM-lig.
- Experimentally/clinically identified and validated mutations that highly destabilize rifampin binding (at wildtype H451 & P489 positions) [9,10].
- Predicted extremely detrimental mutations for protein stability, protein–protein and protein-nucleic affinities (at wildtype K884 & H1035 positions).

3.8. Detrimental mutations in the rifampin binding site:

We have noted that any mutation at rifampin-interacting residues S437, H451, R454, S456, L458, G459, R465, P489, P492 and N493 destabilize protein ligand affinity (mCSM-lig). Of these we have chosen wild-type residues H451 and P489, which are experimentally identified mutations, and wild-type residues S437 and G459, which are computationally predicted (only one mutation was experimentally identified at residue position S437L (reported by us earlier [4], and has destabilizing effects on the overall stability of the protein and affinity to rifampin).

3.9. S437

Serine at position 437 in the wild-type structure forms mainchain and sidechain hydrogen bonds with S434, G432 and R173. The residue has a network of proximal polar interactions and hence stabilizes the rifampin-binding pocket. It was noted that any mutation at this position reduces rifampin affinity (mCSM-lig) and stability of the β-subunit (mCSM) (Supplementary Table 1) (Fig. 7A). The maximum destabilizing effect was noted for substitution to histidine (−1.701 kcal/mol (mCSM)) and it forms hydrogen bonds with S434 and Q438, aromatic bonds with F431, and many ring-ring and π interactions with the surrounding residues which might largely effect the shape of the binding pocket (Fig. 7B). Substitution with leucine causes a minimal destabilizing effect (−0.072 kcal/mol (mCSM)) and stability effects of all the other amino acid substitutions range from −0.072 to −1.701 kcal/mol (mCSM).

S437 is located at 3.3 Å from the interface of β and β′ subunits. Arginine substitution destabilized the interface with the predicted stability change of −0.820 kcal/mol (mCSM-ppi). In the wild-type structure, S437 is located 11.9 Å from the closest nucleic acid molecule but is present on the helix that interacts with both DNA and transcribing RNA in the active center cleft. An aspartate substitution destabilized the protein-RNA interaction with predicted affinity change of −3.857 kcal/mol (mCSM-NA2). S437 is located 4.0 Å from rifampin and forms only proximal interactions with rifampin. However, this residue forms hydrogen bond interactions with S434 and R173 that are important for the attachment of rifampin to the binding pocket. The S437R mutation disrupts the hydrogen bonds with S434 and R173 which in-turn impact stability of rifampin in the binding pocket (−1.062 kcal/mol (mCSM-lig)).

3.10. G459

Glycine at position 459 forms hydrogen bonds with Q435, L458 and G462, and carbonyl interactions with the P460. G459 is present 4.6 Å away from rifampin and is involved in hydrogen bonds with residues that interact with rifampin (Fig. 7C). A tryptophan substitution largely destabilizes the binding pocket by the incorporation of hydrophobic and π interactions with the surrounding residues. It forms side-chain hydrophobic interactions with L436, L384 and F430. It also forms a ring–ring interaction with F430, an atom-ring interaction with L384 and intergroup interactions with Q178 and Q388. It forms multiple hydrogen bonds with the surrounding residues, which may impact the orientation of the binding pocket and destabilize the protein (Fig. 7D).

3.11. Clinically identified mutations that highly destabilize rifampin binding:

From the 40 mutations that are reported from different rifampin-resistant leprosy clinical isolates (Supplementary Table 2), we have chosen two residues where mutations are extremely detrimental to protein stability, protein ligand affinity, protein nucleic affinity and protein subunit interfaces. These substitutions at positions H451 and P489 were studied in detail.

3.12. H451:

H451 in the wild-type structure lies 3.7 Å from rifampin and 4.1 Å from the interface. This residue forms cation – π interactions with guanidinium group of R454, which in turn forms polar interactions with rifampin (Fig. 8A). Additionally, H451 makes two hydrogen bonds with mainchain amino group of R454 and oxygen atom of S447. Mutations at this residue site largely impact the stability and ligand binding. Substitution to serine induced a change in stability of the protein with a decrease in Gibbs free energy of −1.898 kcal/mol and a network of π interactions that are present in the native structure, were lost in the mutant (Fig. 8B).

Methionine substitution destabilizes β – β′ subunit interface and leads to a change in free energy of −1.451 Kcal/mol. Methionine forms carbonyl interactions with K452 and T450, a hydrophobic interaction with Q438 and weak hydrogen bond interactions

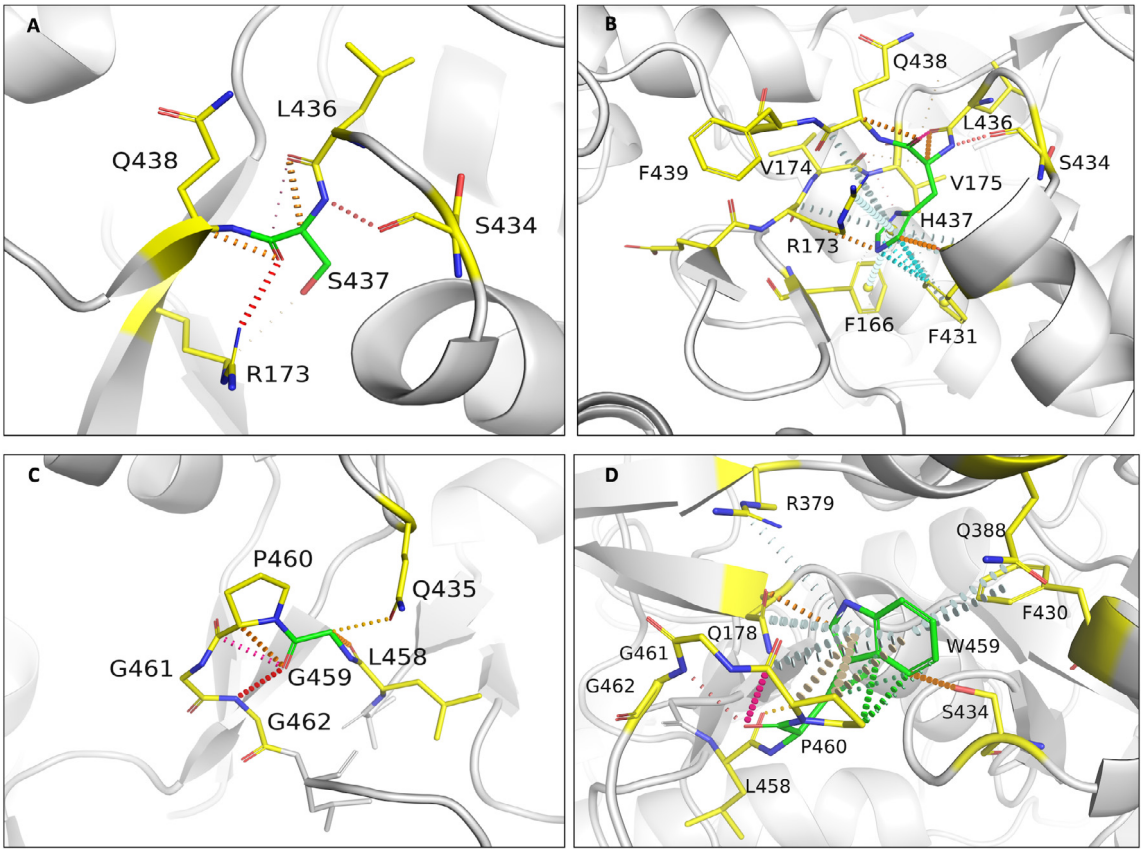


Fig. 7. [A] Interactions of S437 with the surrounding residue environment in the wildtype and of H437 in the S437H mutant [B]. [C] Interactions of G459 with the surrounding residue environment and [D] W459 in the mutant G459W. The red dotted lines represent hydrogen bonds. Orange dotted lines represent weak hydrogen bond interactions. Ring-Ring and intergroup interactions are depicted in cyan. Aromatic interactions are represented in sky-blue and carbonyl interactions in pink dotted lines. Green dotted lines represent hydrophobic interactions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with rifampin. Although histidine or methionine do not directly interact with the residues of the β′ subunit, the changes in the network of π-interactions coupled with the addition of hydrophobic bonds among proximal residues in the interface may change their binding patterns leading to destabilization of the interface.

Substitution with glutamic acid induces a destabilizing effect on the β-subunit-rifampin interaction. E451 forms weak hydrogen bond, carbonyl and proximal hydrophobic interactions with the residue environment but does not form any bonds with rifampin, unlike the wild-type residue that forms proximal hydrogen bonds with rifampin.

3.13. P489

Proline at position 489 is present in a loop which is in close proximity to rifampin and forms hydrophobic interaction with rifampin and weak hydrogen bond interactions with T488 and Q490 (Fig. 8C). Mutations at the position 489 were reported in rifampin-resistant leprosy patients from Thailand [9]. Glycine substitution destabilizes the protein (−1.771 kcal/mol) leading to a loss of hydrophobic interaction with rifampin. Weak hydrogen bond and carbonyl interactions, however, were retained in the mutant model (Fig. 8D). Arginine substitution destabilizes interface and rifampin affinities, with predicted stability changes of −1.372 and −0.917 kcal/mol respectively. FoldX predicted a large change in stability of 4.79 kcal/mol for difference between mutant and wild types, which is highly destabilizing. FoldX optimizes the sidechains and moves the structure to a lowest energy state (usually represented as a negative value) and hence the difference between two negative energy values of wild and mutant is considered destabilizing.

3.14. Extremely detrimental mutations:

Mutations at residues positions K884 and H1035 were considered to be extremely detrimental. These residues lie in close proximity to the interface, nucleic acids and rifampin. Substitutions at these sites destabilize protomer, protein–protein interfaces (both the residues reside at the subunit interface), protein-nucleic acid and protein–ligand affinities. Both empirical (FoldX) and knowledge based (mCSM and SDM) methods predicted destabilizing effects.

3.15. K884

K884 is located 3.2 Å from the interface, 3.3 Å from the nucleic acid and 8.6 Å from rifampin. Lysine forms mainchain hydrogen bonds with L1033 and proximal hydrophobic interactions with H1035 and V894. It also forms a cation - π interaction with H1035 and most importantly a sidechain proximal hydrogen bond with the sugar phosphate group of guanine (second) nucleotide in the RNA transcript. This interaction is crucial for maintaining the RNA interaction with rifampin in order to induce steric clash on the adjacent nucleotide and halt transcription (Fig. 9A). Serine substitution at this site results in the loss of this vital interaction. S884 forms weak Van der Waals interactions with D883 and L885 and

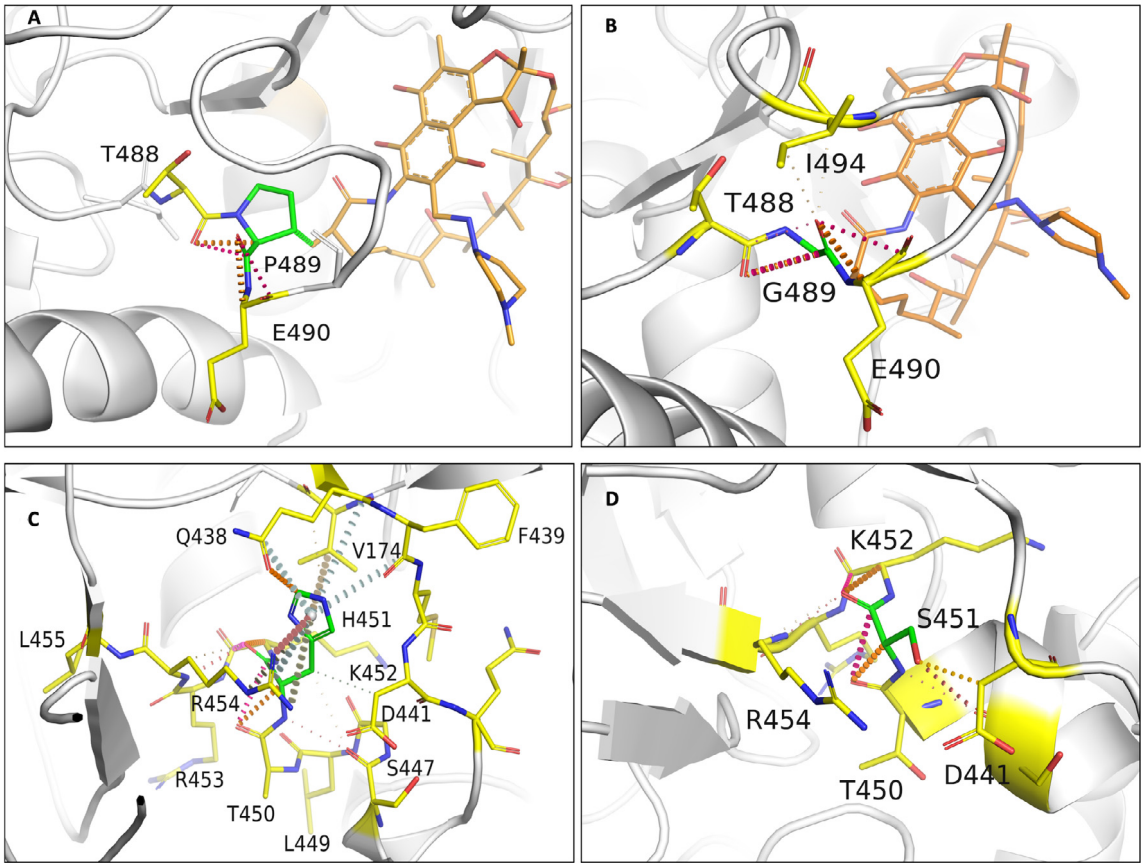


Fig. 8. [A] Interactions of P489 with the surrounding residue environment in the wildtype and of G489 in the P489G mutant [B]. [C] Interactions of H451 with the surrounding residue environment and [D] S451 in the mutant H451S.

hydrogen bonds with L1033 and H1035. Interactions with RNA backbone are lost in the mutant (Fig. 9B). The mutant is destabilized (−2.298 kcal/mol).

Aspartate substitution at this site destabilizes RNA affinity (−2.130 kcal/mol) and the mutant residue forms hydrogen bonds with L1033 and H1035, and hydrophobic interactions with V894.

3.16. H1035

Histidine at position 1035 is located 3.5 Å from the interface and RNA, and 8.8 Å away from rifampin. It forms a network of π interactions with the surrounding residues. The ring-ring π interactions with the fused pyrimidine-imidazole ring of guanine in the first nucleotide of RNA transcript is vital to the orientation of RNA transcript in the active center cleft (Fig. 9C). These interactions are lost in substitutions with non-aromatic amino acids. It was also noted that aspartate substitution largely destabilizes β subunit -rifampin affinity (Fig. 9D).

3.17. Impact of mutations on flexible conformations:

The stability changes between the wildtype and each mutant in lowest energy conformation were calculated by FoldX and have a Pearson's correlation coefficient ("r" value) of 0.38 with other predictors mCSM and SDM. Although FoldX does not probe backbone conformational changes, it optimizes the sidechain rotamers of the mutant residues to attain a low energy state and calculates the change in free energy between the states. We further sampled the fully flexible conformers of the β-subunit and estimated changes in vibrational entropy ΔS and protein stability using ENCoM. A linear combination of vibrational entropy ΔS by ENCoM and enthalpy changes by FoldX were used to calculate stability changes. ENCoM predicted highly destabilizing mutations in the rifampin binding and RNA interacting sites in the active center cleft of the holoenzyme. DynaMut predictions correlated with ENCoM values at an r value of 0.56. The average change in stability predicted by ENCoM and DynaMut for any mutation at each residue position in the β subunit was mapped on the model (Fig. 10A and B).

3.18. Protein stability changes and fragment hotspot maps:

Fragment hotspots were mapped on the structure that is colored by regions predicted to have least protein stability changes due to any mutations (using mCSM, SDM and FoldX software). As fragment hotspot maps program identifies small molecule binding propensity on the surface of the protein, we used only the protein stability prediction software to identify areas that are stable by any mutations. The regions of the β subunit that are least impacted by mutations (mutation coolspots) are overlaid with fragment hotspot maps. The site B (Fig. 11), which is in close proximity to the RNA binding region and is a pocket at the β-β′ subunit interface, is least impacted by mutations and has a hotspot at the contouring score of 17 with donor, apolar and acceptor regions [22]. Secondly, the site A, although located away from the catalytic core of the enzyme, is present in the path of entry/exit point for template DNA into the holoenzyme complex and a small molecule interaction at this site can potentially impact template DNA interactions or induce conformational change in the crab-claw-shaped β subunit leading to disruption in the holoenzyme assembly.

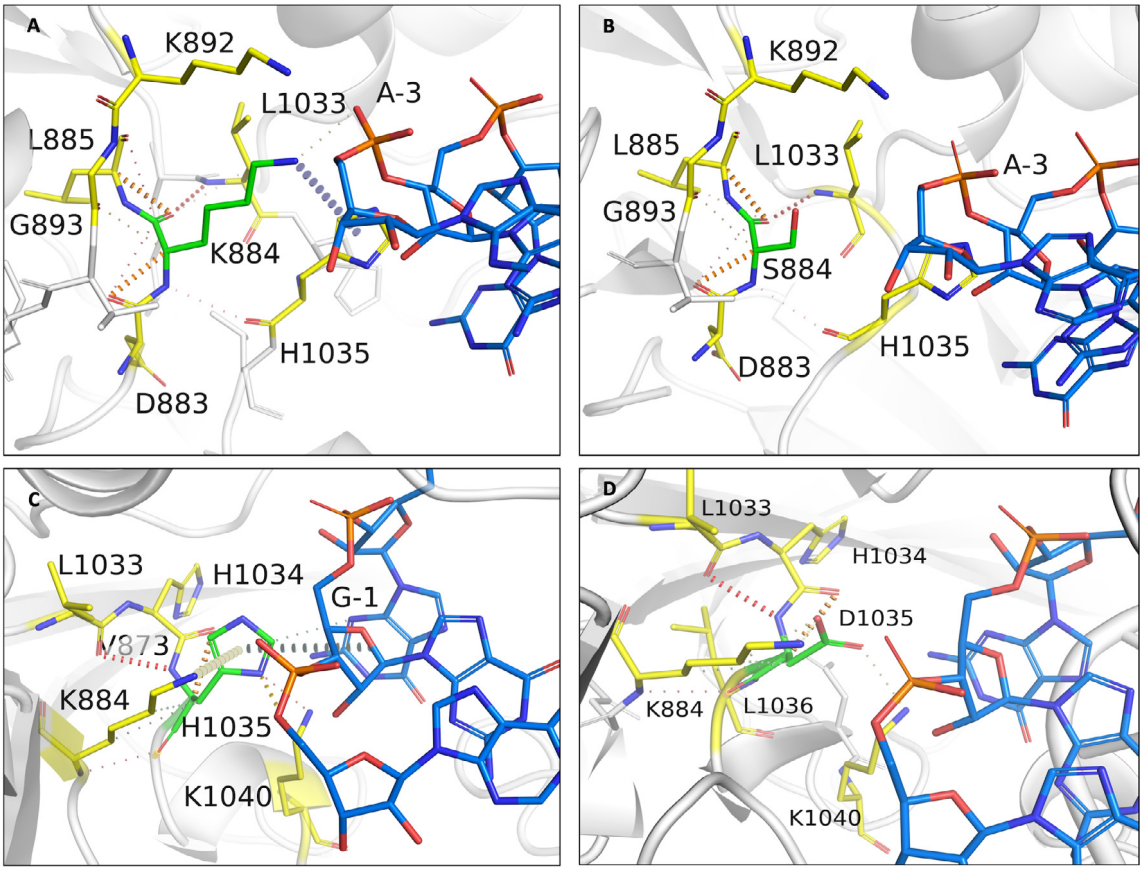


Fig. 9. [A] Interactions of K884 with the surrounding residue environment in the wildtype and of S884 in the K884S mutant [B]. [C] Interactions of H1035 with the surrounding residue environment and [D] D1035 in the mutant H1035D. The blue dotted lines represent cation-π interaction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

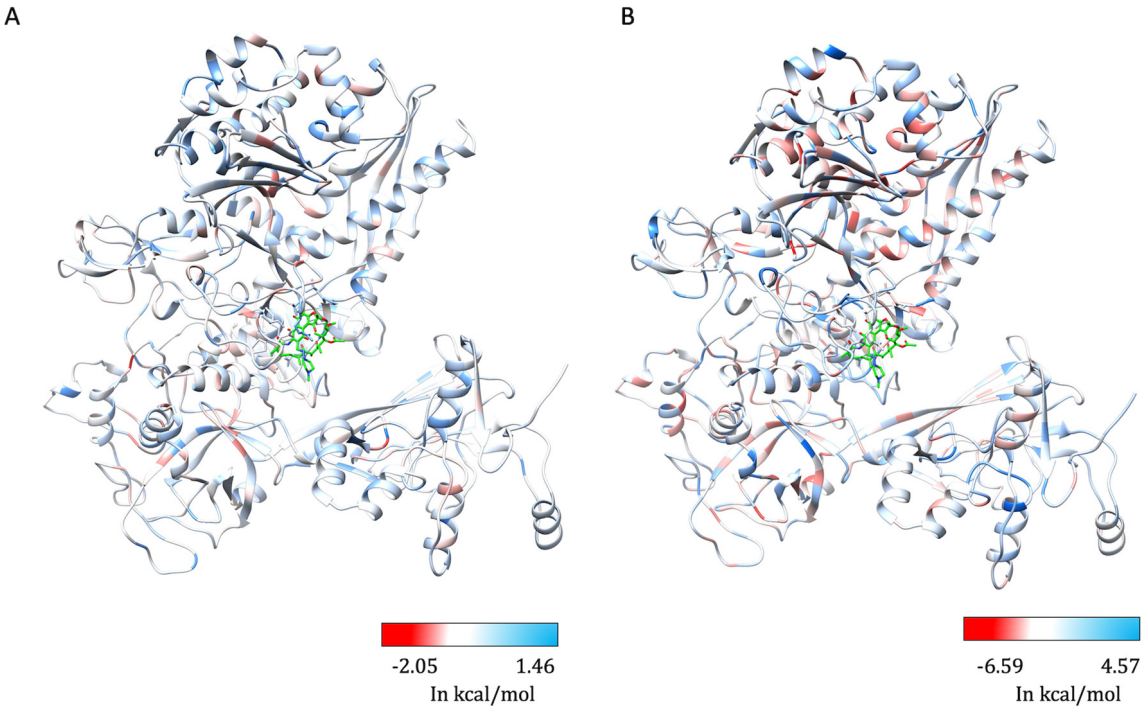


Fig. 10. [A] The maximum destabilizing effects on the protein stability, a mutation can induce at each residue position in the flexible conformations (as predicted by ENCoM [A] and DynaMut [B]), are mapped on the structure. Regions in red represent highly destabilizing while the blue regions are relatively stable with mutations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

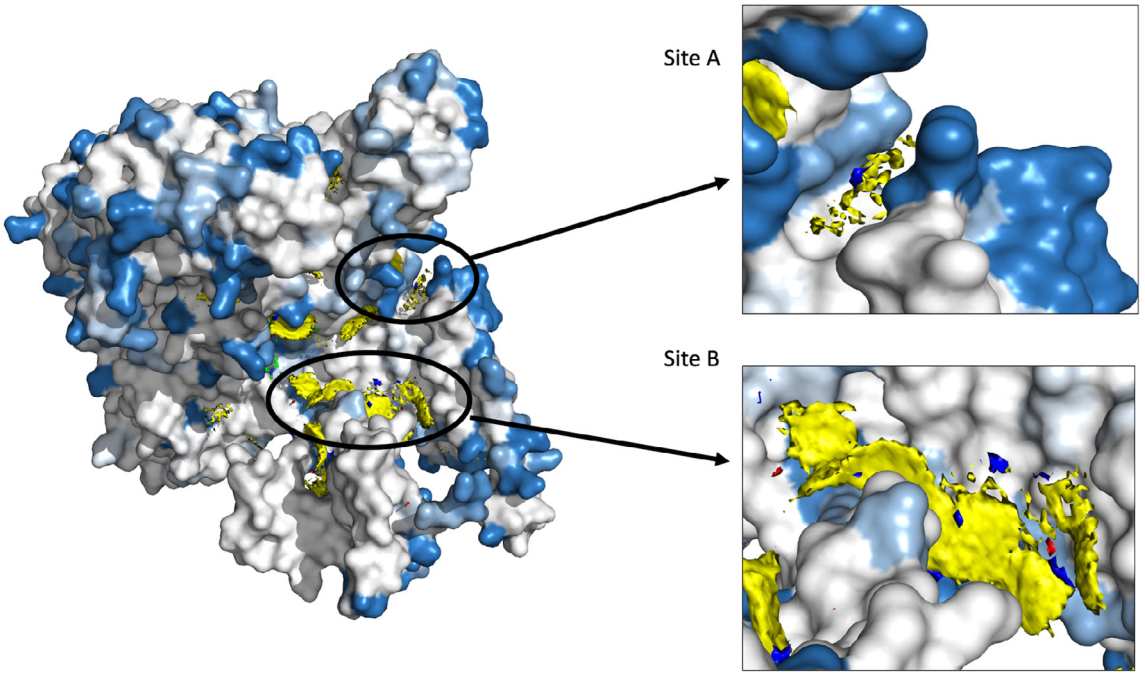


Fig. 11. Fragment hotspots were mapped on the structure which was coloured with maximum destabilizing effects of systematic mutations at each residue positions. Blue represents regions which are least impacted by any mutations. Stable and potential small molecule binding sites "A" and "B" are depicted on the structure.

4. Discussion

In the absence of a rapid and an effective laboratory-based diagnostic tool for determining drug resistance in leprosy, identification of mutations known to confer resistance to individual drugs in MDT remains an appropriate approach for diagnosing drug resistance. Associations between mutations in the drug targets and clinical resistance to individual drugs in MDT are often validated by mouse-footpad experiments in which, resistant strains (with known mutations) are propagated in the hind footpads of mice (cross-bred albino) in the presence of drugs under study [4]. Owing to high percentage identity of the β subunit of RNAP of *M. leprae* with that of *M. tuberculosis,* identical mutations that are experimentally proven to confer rifampin resistance in tuberculosis, are considered as likely drug-resistant mutations in leprosy. The experimentally known mutations in *M. leprae* were those identified by DNA sequencing of *rpoB* gene (derived from skin tissue DNA of relapsed/drug resistant leprosy patients) and published in different studies (reference for each mutation is listed in Supplementary Table 2). Most of these were validated in either mouse foot-pad experiments or by using surrogate genetic hosts [5].

Around 40 different rifampin-resistance mutations were noted in *M. leprae* from clinical isolates around the world using amplicon sequencing of RRDR [10]. All of these mutations decrease the stability of rifampin binding to the β-subunit (Supplementary Table 2) and the mutant strains exhibited normal grown patterns in the mouse footpads when administered with rifampin in doses equivalent to WHO regimen of multibacillary MDT [44]. This indicates that mutations structurally and functionally impact rifampin interactions and influence concomitant resistance.

Thermodynamic stability of the proteins essentially influences their function and is largely dependent on the sequence. Missense mutations that lead to amino acid substitutions often impact protein stability, shifting it towards either a stabilized or a destabilized state [7]. Experimental measurements of stability changes in proteins are often challenging especially with large and complex protein machineries like RNAP. However, mutations within each subunit of the RNAP complex, and primarily the rifampin binding β-subunit, have clinical implications and influence rifampin-resistance outcomes in mycobacterial diseases [45]. The performance of various structural, sequence and NMA based predictors for predicting protein stability changes upon mutations vary largely in terms of their accuracy and bias [46], but offer a quick and a helpful alternative to understanding the association between mutations and resistance phenotypes [6].

Given the absence of a rapid and experimentally validated system to read the impact of mutations in the β-subunit of RNAP in *M. leprae* with clinical rifampin resistance outcomes in leprosy, we conducted computational saturation mutagenesis to determine regions on the β-subunit that impact the overall stability, protein-subunit interfaces, protein-nucleic and protein–ligand affinities. Being a part of the complex transcriptional machinery in the mycobacterial cell, the compositional and conformational stability of the β-subunit is crucial to binding of DNA template and synthesis of complementary RNA transcript in the active center cleft of the holoenzyme [47,48]. As rifampin blocks the growing RNA transcript through steric occlusion, its binding and orientation in the binding pocket is vital to its function [47]. Mutations within the RRDR impact rifampin interactions and overall stability of the subunit. As noted from Supplementary Table 2, all the experimentally identified *rpoB* gene mutations from *M. leprae* indicated a destabilizing effect on the protein–ligand affinity. Owing to the robustness of these predictions, we employed an *in-silico* saturation mutagenesis model to understand the impacts of systematic mutations at each residue site of the subunit.

The destabilizing mutations are given preference over mutations that are silent or have minimal effects on the stability. This is to explore and understand the possible structural and functional implications of emerging detrimental mutations (reported or new) that can influence rifampin resistance outcomes in leprosy. We used different structural, sequence and NMA based tools to identify and compare the predictions. mCSM stability predictions had better correlations with the other predictors (SDM (r = 0.55),

MAESTRO (r = 0.61), Imutant 2.0 Structure (r = 0.72), CUPSAT (r = 0.43), Imutant 2.0 Sequence (r = 0.62) and DynaMut (r = 0.61)).

Protocols (Computational Saturation Mutagenesis (CoSM)) [49] that use molecular dynamic equilibration, sidechain flips and energy minimization to improve side conformations in mutants enable prediction of stability changes with better accuracy and correlation with the experimentally deciphered stability changes (r = 0.9). However, these protocols are computationally intensive and require high performance computing systems and time. CoSM had a similar performance to FoldX, which was used in the current study. Given the large sample size, molecular dynamic equilibration of sidechain rotamers is beyond the scope of this study.

In conclusion, we have deciphered the predicted effects of all possible mutations in the β-subunit of RNAP in *M. leprae* using computational saturation mutagenesis model, probing structural, sequence driven and dynamic changes that impact overall stability of the protein, RNA and rifampin affinities. The predicted impacts were mapped onto the structures and highly detrimental mutations were further analyzed for their changes in interatomic interactions. Due to the lack of adequate experimental data on stability changes in β-subunit of RNAP upon mutations, we have limited information on the accuracy of the predictions, however, all the prediction tools used in the study are well tested and validated software which are proven to perform with reasonable accuracy and minimal bias on various relevant mutational datasets [31]. To date there were no studies describing the phenotypic resistance/susceptibility outcomes in strains with compensatory mutations in RNAP. Further studies on saturation mutagenesis of the entire RNAP holoenzyme complex may provide comprehensive information on the effects of co-evolving and compensatory mutations in other subunits on rifampin binding and function.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors would like to thank the rest of the computational biology team at Department of Biochemistry, University of Cambridge, United Kingdom, for their overarching support and guidance in the data collection and analysis. SCV was supported by American Leprosy Missions, United States of America, (Grant No: G88726), MJS was supported by a grant from Foundation Botnar working to support children with cystic fibrosis, Switzerland (Project 6063), CHMR and SP were supported by Australian Government Research Training Program Scholarships, Australia, DBA was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council, United Kingdom and Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Brazil (MR/M026302/1) and by the Wellcome Trust Programme Grant, United Kingdom (200814/Z/16/Z) and supported in part by the Victorian Government's OIS Program, Australia. TLB was supported by the Wellcome Trust Programme Grant, United Kingdom (200814/Z/16/Z) and SM was supported by the MRC DBT Grant, United Kingdom and India (RG78439).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.01.002.

References

[1] World Health Organization. Global tuberculosis report 2018 [Internet]. World Health Organization; 2018 [cited 2018 Dec 19]. 231 p. Available from: http://apps.who.int/iris/handle/10665/274453
[2] Han XY, Sizer KC, Thompson EJ, Kabanja J, Li J, Hu P, et al. Comparative sequence analysis of *Mycobacterium leprae* and the new leprosy-causing *Mycobacterium lepromatosis*. J Bacteriol 2009;191(19):6067–74.
[3] WHO | Weekly Epidemiological Record, 31 August 2018, vol. 93, 35 (pp. 444–456) [Internet]. WHO. [cited 2018 Dec 19]. Available from: http://www.who.int/wer/2018/wer9335/en/.
[4] Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, et al. Structural implications of mutations conferring rifampin resistance in *Mycobacterium leprae*. Sci Rep 2018;8(1):5016.
[5] Nakata N, Kai M, Makino M. Mutation analysis of mycobacterial rpoB genes and rifampin resistance using recombinant *Mycobacterium smegmatis*. Antimicrob Agents Chemother 2012;56(4):2008–13.
[6] Pires DEV, Chen J, Blundell TL, Ascher DB. *In silico* functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 2016;22(6):19848.
[7] Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N. Understanding molecular consequences of putative drug resistant mutations in *Mycobacterium tuberculosis*. Sci Rep 2018;8(1):15356.
[8] Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, et al. Analysis of a novel *pncA* mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 2018;198(4):541–4.
[9] Ramasoota P, Wongwit W, Sampunachot P, Unnarat K, Ngamying M, Svenson SB. Multiple mutations in the rpoB gene of *Mycobacterium leprae* strains from leprosy patients in Thailand. Southeast Asian J Trop Med Public Health 2000;31(3):493–7.
[10] Cambau E, Saunderson P, Matsuoka M, Cole ST, Kai M, Suffys P, et al. Antimicrobial resistance in leprosy: results of the first prospective open survey conducted by a WHO surveillance network for the period 2009–15. Clin Microbiol Infect 2018;24(12):1305–10.
[11] Williams DL, Gillis TP. Drug-resistant leprosy: monitoring and current status. Lepr Rev 2012;83(3):269–81.
[12] Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med 2009;6(2):e1000002.
[13] Htike Min PK, Pitaksajjakul P, Tipkrua N, Wongwit W, Jintaridh P, Ramasoota P. Novel mutation detection IN rpoB OF rifampicin-resistant *Mycobacterium tuberculosis* using pyrosequencing. Southeast Asian J Trop Med Public Health 2014;45(4):843–52.
[14] André E, Goeminne L, Colmant A, Beckert P, Niemann S, Delmee M. Novel rapid PCR for the detection of Ile491Phe rpoB mutation of *Mycobacterium tuberculosis*, a rifampicin-resistance-conferring mutation undetected by commercial assays. Clin Microbiol Infect 2017;23(4):267.e5–7.
[15] Al-Mutairi NM, Ahmad S, Mokaddas E, Eldeen HS, Joseph S. Occurrence of disputed rpoB mutations among *Mycobacterium tuberculosis* isolates phenotypically susceptible to rifampicin in a country with a low incidence of multidrug-resistant tuberculosis. BMC Infect Dis 2019;19(1):3.
[16] Lahiri N, Shah RR, Layre E, Young D, Ford C, Murray MB, et al. Rifampin resistance mutations are associated with broad chemical remodeling of Mycobacterium tuberculosis. J Biol Chem 2016;291(27):14248–56.
[17] Andres S, Gröschel MI, Hillemann D, Merker M, Niemann S, Kranzer K. A diagnostic algorithm to investigate pyrazinamide and ethambutol resistance in rifampin-resistant *Mycobacterium tuberculosis* isolates in a low-incidence setting. Antimicrob Agents Chemother 2019;63(2):e01798–e1818.
[18] Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 2017.
[19] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res 2005;33:W382–8. https://doi.org/10.1093/nar/gki387.
[20] Frappier Vincent, Chartier Matthieu, Najmanovich Rafael J. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Res 2015;43:W395–400. https://doi.org/10.1093/nar/gkv343.
[21] Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46(W1):W350–5.
[22] Radoux CJ, Olsson TSG, Pitt WR, Groom CR, Blundell TL. Identifying interactions that determine fragment binding at protein hotspots. J Med Chem 2016;59(9):4314–25.
[23] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234(3):779–815.
[24] Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 2004;32:W615–9. https://doi.org/10.1093/nar/gkh398.
[25] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 2004;25(13):1605–12.
[26] Smith RE, Lovell SC, Burke DF, Montalvao RW, Blundell TL. Andante: reducing side-chain rotamer search space during comparative modeling using

environment-specific substitution probabilities. Bioinformatics 2007;23(9):1099–105.
[27] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9.
[28] Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30(3):335–42.
[29] Pires DEV, Ascher DB. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 2017.
[30] Pires DEV, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 2016;7(6):srep29575.
[31] Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO – multi agent stability prediction upon point mutations. BMC Bioinf 2015;16(1):116.
[32] Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 2006;34(Web Server issue):W239–42. https://doi.org/10.1093/nar/gkl190.
[33] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33(Web Server issue):W306–10. https://doi.org/10.1093/nar/gki375.
[34] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012;7(10):e46688.
[35] Ashkenazy Haim, Abadi Shiran, Martz Eric, Chay Ofer, Mayrose Itay, Pupko Tal, Ben-Tal Nir. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 2016;44(W1):W344–50. https://doi.org/10.1093/nar/gkw408.
[36] Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 2017;429(3):365–71.
[38] Strub C, Alies C, Lougarre A, Ladurantie C, Czaplicki J, Fournier D. Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. BMC Biochem 2004;5:9.
[39] Wilson KA, Holland DJ, Wetmore SD. Topology of RNA–protein nucleobase–amino acid π–π interactions and comparison to analogous DNA–protein π–π contacts. RNA 2016;22(5):696–708.
[40] Dougherty DA. Cation-π interactions involving aromatic amino acids. J Nutr 2007;137(6):1504S–8S.
[41] Gallivan JP, Dougherty DA. Cation-π interactions in structural biology. PNAS 1999;96(17):9459–64.
[42] Jamieson FB, Guthrie JL, Neemuchwala A, Lastovetska O, Melano RG, Mehaffy C. Profiling of rpoB mutations and MICs for rifampin and rifabutin in *Mycobacterium tuberculosis*. J Clin Microbiol 2014;52(6):2157–62.
[43] Miotto P, Cabibbe AM, Borroni E, Degano M, Cirillo DM. Role of disputed mutations in the rpoB gene in interpretation of automated liquid mgit culture results for rifampin susceptibility testing of *Mycobacterium tuberculosis*. J Clin Microbiol 2018;56(5).
[44] Colston MJ, Hilson GR, Banerjee DK. The "proportional bactericidal test": a method for assessing bactericidal activity in drugs against *Mycobacterium leprae* in mice. Lepr Rev 1978;49(1):7–15.
[45] Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 2012;44(1):106–10.
[46] Usmanova DR, Bogatyreva NS, Ariño Bernad J, Eremina AA, Gorshkova AA, Kanevskiy GM, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Bioinformatics 2018;34(21):3653–8.
[47] Lin W, Mandal S, Degen D, Liu Y, Ebright Y, Li S, et al. Structural basis of Mycobacterium tuberculosis transcription and transcription inhibition. bioRxiv 2017:099606.
[48] Boyaci H, Chen J, Lilic M, Palka M, Mooney RA, Landick R, et al. Fidaxomicin jams *Mycobacterium tuberculosis* RNA polymerase motions needed for initiation via RbpA contacts. Elife 2018;26:7.
[49] Fischer A, Seitz T, Lochner A, Sterner R, Merkl R, Bocola M. A fast and precise approach for computational saturation mutagenesis and its experimental validation by using an artificial (βα)8-barrel protein. ChemBioChem 2011;12(10):1544–50.

Appendix M

Structure guided prediction of pyrazinamide resistance mutations in pncA

SCIENTIFIC REPORTS
nature research

OPEN

Structure guided prediction of Pyrazinamide resistance mutations in pncA

Malancha Karmakar[1,2,3], Carlos H. M. Rodrigues[1,2], Kristy Horan[4], Justin T. Denholm[3] & David B. Ascher[1,2,5*]

Pyrazinamide plays an important role in tuberculosis treatment; however, its use is complicated by side-effects and challenges with reliable drug susceptibility testing. Resistance to pyrazinamide is largely driven by mutations in pyrazinamidase (pncA), responsible for drug activation, but genetic heterogeneity has hindered development of a molecular diagnostic test. We proposed to use information on how variants were likely to affect the 3D structure of pncA to identify variants likely to lead to pyrazinamide resistance. We curated 610 pncA mutations with high confidence experimental and clinical information on pyrazinamide susceptibility. The molecular consequences of each mutation on protein stability, conformation, and interactions were computationally assessed using our comprehensive suite of graph-based signature methods, mCSM. The molecular consequences of the variants were used to train a classifier with an accuracy of 80%. Our model was tested against internationally curated clinical datasets, achieving up to 85% accuracy. Screening of 600 Victorian clinical isolates identified a set of previously unreported variants, which our model had a 71% agreement with drug susceptibility testing. Here, we have shown the 3D structure of pncA can be used to accurately identify pyrazinamide resistance mutations. SUSPECT-PZA is freely available at: http://biosig.unimelb.edu.au/suspect_pza/.

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, is the leading cause of infectious disease death worldwide. In 2017, 10 million people fell ill, and 1.6 million died, from tuberculosis[1]. While a range of antibiotics are available to treat TB, treatment is prolonged, and the increasing emergence of drug-resistant bacteria is a considerable threat to global health. In 2017 alone, an estimated 558,000 people developed multi-drug-resistant tuberculosis (MDR-TB), resistant to the two first-line drugs rifampicin and isoniazid[1].

Pyrazinamide (PZA) is a first-line drug that exhibits unique sterilizing activity towards both drug-susceptible and MDR-TB[2]. It is responsible for the killing of the persistent tubercle bacilli during the initial intensive phase of chemotherapy, allowing treatment to be shortened from 9 months to 6 months for drug susceptible cases[3]. PZA therapy has been linked to improved outcomes for both non-MDR and MDR-TB, and is being considered as part of the future regimens in combinations with bedaquiline, delamanid, PA-824 and moxifloxacin, which are currently in phase three trials[4,5].

Despite the highly important role of PZA in clinical outcomes, resistance has largely been underestimated, with up to 20% of non-MDR-TB patients PZA resistant[6]. Being a central drug in current and future regimens, it is important to be able to rapidly and accurately identify resistant isolates and track the emergence and spread of drug resistant strains. *In vitro* drug susceptibility testing (DST) is challenging, expensive and time-consuming as PZA is effective against *M. tuberculosis* only at acidic pH, leading to false resistance rates of up to 70%[7–13]. This has led to the WHO recommending the development of molecular genetics tests.

PZA is a structural analog of nicotinamide and is a pro-drug that needs to be converted into its active form, pyrazinoic acid (POA), by the non-essential enzyme pyrazinamidase, encoded by the *pncA* gene[14,15]. It has been

[1]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia. [2]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia. [3]Victorian Tuberculosis Program, Melbourne Health and Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia. [4]Microbiological Diagnostic Unit Public Health Laboratory, University of Melbourne at The Peter Doherty Institute for Infection &Immunity, Melbourne, Victoria, Australia. [5]Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK. *email: david.ascher@unimelb.edu.au

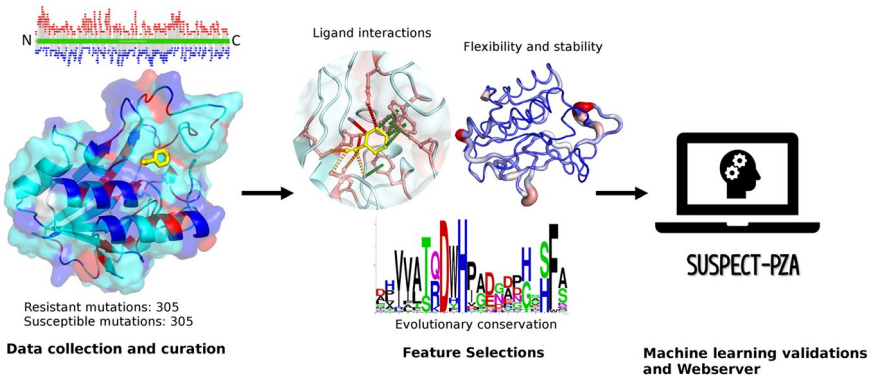


Figure 1. Methodology workflow. The methodology can be divided into three steps. In step 1, data is collected and curated from various tuberculosis databases and articles with experimental evidence like availability of DST results or high-precision laboratory screening study. The curated mutations are shown across both the protein sequence and 3D structure, respectively. The protein sequence and structure of PncA is colored by whether resistant (red) or susceptible (blue) mutations have been observed at that location. Highlighting the difficulty of genomic analysis of pncA, both resistant and susceptible mutations have been observed across many residue positions (cyan). In step 2, effects of mutations on protein stability, dynamics, complementary information regarding the environment characteristics of the wild-type residue (e.g. relative solvent accessibility, residue depth and secondary structure), PZA binding affinity are calculated using different *in-silico* tools. Step 3, all the features are used as evidence to train a supervised machine learning algorithm and after evaluating the performance of the predictive model, the consensus predictions are integrated into a server and can be used to guide clinical resistance detection.

postulated that the mechanism of action of PZA is through POA, which disrupts the bacterial membrane energetics and inhibits the membrane transport function which is necessary for the survival of the bacterium, at an acidic site of infection[16]. PZA resistance has been linked to mutations in a number of genes, including *pncA*, *rpsA*[17], *panD*[18], *clpC1*[19], and the putative efflux pumps *Rv0191*, *Rv3756c*, *Rv3008*, and *Rv1667c*[20], but mutations in *pncA* are the major mechanism for PZA resistance (70–97%)[21]. While sequencing the *pncA* gene can be a more reliable method to determine resistance than DST, which is prone to missing low-level pyrazinamide resistance caused by non-synonymous mutations in pncA[22], the development of a genetics based resistance screen is complicated as resistant and non-resistant mutations are found across the entire protein.

To solve the problem of a reliable DST for PZA, we previously showed that protein structural information can be used in a clinical setting to rapidly, accurately and pre-emptively predict drug resistant mutations in *pncA*[23]. This showed that mutations that affected protein folding, flexibility, stability and activity were strongly associated with resistance. Here we have used a comprehensive combination of structure and sequence-based features to develop a predictive tool to characterize novel PncA mutations, which we tested on novel mutations from the Victorian Tuberculosis Program, CRyPTIC[24] and Miotto *et al.* dataset[25]. This highlights the potential of using structural information to guide the genetic detection of resistance. We have implemented our model through the webserver SUSPECT-PZA (http://biosig.unimelb.edu.au/suspect_pza/), which will enable the rapid structural evaluation of the molecular and phenotypic consequences of any *pncA* nonsynonymous mutation to support informed clinical decisions.

Results

We used a structure-guided approach to understand the structural and functional consequences of variants in the drug target PncA, and machine learning to build an empirical tool that could identify likely resistant mutations. The workflow used to analyze the mutations and train a Random Forest algorithm is shown in Fig. 1 and it comprises three major steps: (1) data curation, which can be subdivided into mutational data set acquisition and protein structure curation; (2) feature analysis, which involves the generation and evaluation of features selected to develop the predictive model to determine novel drug resistance mutations in PncA; (3) machine learning and webserver development, which aims to train, test and validate a supervised machine learning algorithm to accurately predict the susceptibility of the variant followed by a database (SUSPECT-PZA) which has information for all possible variants of PncA.

Distribution of the mutations on the structure. We curated a dataset of 1322 nonsynonymous substitutions with high quality experimentally measured PZA susceptibility (71 susceptible mutations from GMTV[26], 12 resistant mutations from GMTV[26], 178 resistant mutations from TBdreamDB[27], Fig. 2A, 547 resistant and 514 susceptible mutations from experimental saturation mutagenesis[28]). After removal of duplicate mutations, we were left with a dataset of 610 mutations, which included 305 susceptible and 305 resistant mutations. Mapping the complete set of curated 610 nsSNVs (Fig. 1) and just the clinical variants only (Fig. 2B) onto the crystal structure of PncA revealed that variants were distributed throughout the entire protein structure, complicating resistance inference from sequence analysis. We also observed that the resistance mutations were not solely localized at the drug binding site but distributed throughout the protein (Fig. 2C).

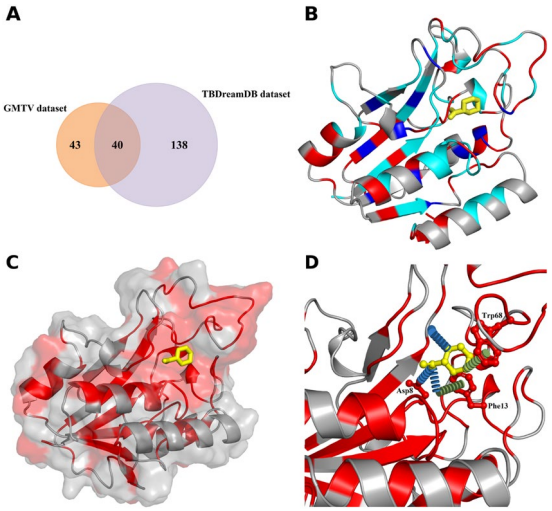


Figure 2. Distribution of clinical resistant and susceptible mutations in PncA. (**A**) Venn diagram representing the distribution of clinical mutations in the different datasets used to build the predictive model. (B) Clinical resistant and susceptible mutations mapped on the crystal structure. Amino acid positions where both susceptible and resistant mutations were seen are colored in cyan and emphasizes the need for a better and improved tool to classify them accurately. (**C**) Surface view of PncA with the docked PZA (yellow, ball and stick representation). Clinical resistant mutations, shown in red, are not just located at the PZA binding site, but are spread equally throughout the whole protein. (**D**) Molecular interactions between PZA (yellow sticks) and the surrounding amino acids which are part of the catalytic triad (Asp8) and substrate binding site (Trp68, Phe13). Hydrogen bonds are shown as blue dashes, and π-interactions as green dashes.

PncA is a small protein molecule which constitutes of 186 amino acids. The experimental crystal structure of the drug (PZA) bound to the enzyme (PncA) was unavailable. Therefore, PZA was *ab* initio docked into the experimental crystal structure of the holo-wild-type PncA protein (PDB ID: 3PL1[29]). The docked structure revealed that PZA formed key interactions within the proteins active site, which includes the catalytic triad (Asp8, Lys96, and Cys138), substrate-binding residues (Trp68 and Phe13), and the iron center (Asp49, His51, His57, and Fe 21). Analysis of the molecular interactions with Arpeggio[30] highlighted a strong network of polar and π- interactions between PZA and PncA (Fig. 2D).

Structural, biophysical and evolutionary consequences of PncA mutations. Looking at the SNAP2[31] and PROVEAN[32] scores, which consider evolutionary information to predict functionally important nonsynonymous mutations, we observed that resistant mutations were always associated with deleterious scores, while susceptible mutations were scored neutral (Table S1; Fig. 3). This suggest that although mutations were spread throughout the protein, mutations associated with resistance were having a stronger effect on the structure and function of the protein.

The wild-type environment also provided information to differentiate between resistant and susceptible mutations, which included relative solvent accessibility (RSA), residue depth and secondary structure of the wild-type residue (Table S1; Fig. 3). This showed that resistant mutations tended to be found at buried residues that were less solvent exposed (average RSA of 0.18 for resistant mutations compared to 0.39 for susceptible; average residue depth of 1.09 Å for resistant mutations compared to 0.75 Å for susceptible; Table S1). These values were consistent with susceptible mutations being in regions that have milder effects on protein stability and activity than the resistance mutations.

The impact of the resistant and susceptible mutations on protein folding, stability and conformation were assessed using biophysical tools which relies on graph-based signatures to calculate the change in Gibb's free energy, like mCSM-Stability[33], DUET[34] and DynaMut[35]. The effect of the mutations on the binding affinity for PZA were assessed using mCSM-Lig[36]. We observed that resistant mutations led to large decreases in PncA stability and conformational flexibility, while susceptible mutations were associated with milder changes (Table S1; Fig. 3). This is consistent with what we have observed previously for non-essential and drug activating proteins[37]. While resistant mutations, however, tended to be located closer to the PZA binding site (average < 10 Å from the PZA; Fig. 3), we did not see a significant difference in the distribution of the effects of resistant and susceptible mutations on PZA binding affinity (Table S1, Fig. S2), likely due to the importance of other molecular effects leading to resistance.

Machine learning to predict PZA resistance. Building on this structural and sequence-based analysis, we tested whether the information generated from these features could be used to train a supervised machine learning algorithm capable of accurately predicting resistant mutations in PncA. We grouped our features into five distinct categories: stability, dynamics, evolutionary conservation, ligand interactions and backbone geometry (structural environment). The performance of predictive models trained on each class of feature was evaluated

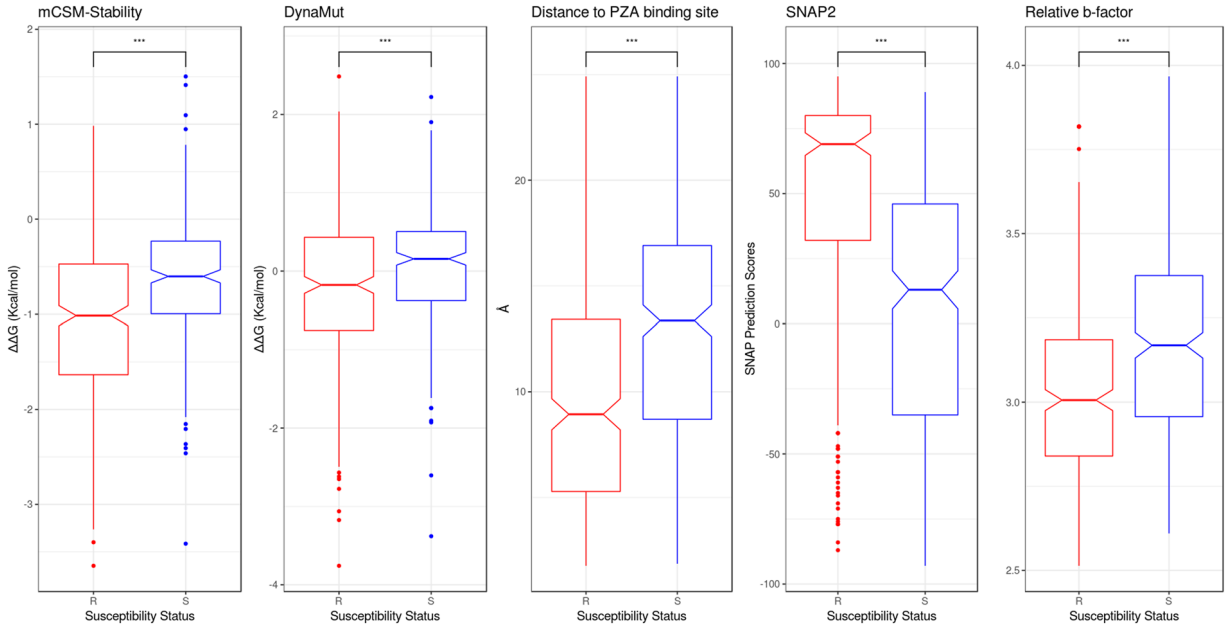


Figure 3. PCA analysis of key molecular features distinguishing resistant and susceptible mutations. Features used for model building are represented as boxplots for explanatory data analysis. The resistant associated mutations (R) are represented as red and the susceptible mutations (S) as blue. (***$p < 0.0001$, Welch two sample t-test).

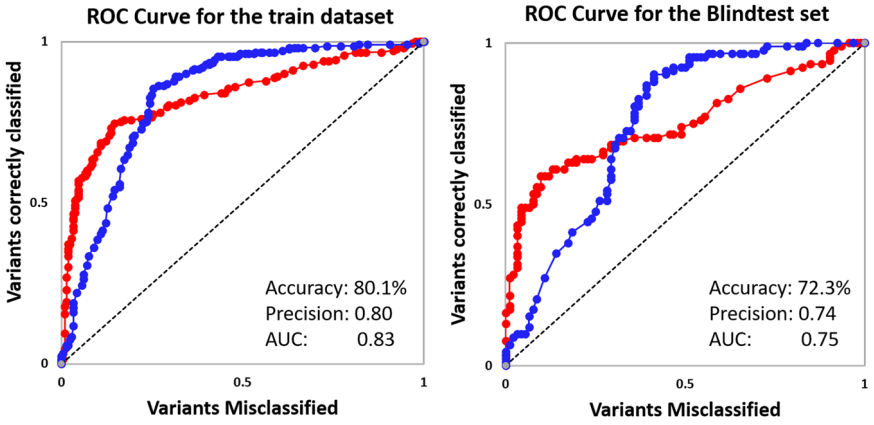


Figure 4. Evaluation Metric for machine learning. Receiver Operating Characteristic (ROC) curves of PZA classifier obtained using the structural and functional consequences of the mutations to accurately identify resistant (red) and susceptible (blue) mutations. (AUC = area under the curve).

separately to explore the contribution of each class to the predictive model (Table S2; Fig. S2). We were able to confirm that the individual categories of features did not yield a good metric for a reliable predictive model, but in combination using 10-fold cross-validation, models trained using Random Forest algorithm yielded a more balanced and accurate performance, highlighting the synergistic effect of these features. The final model correctly classified 80.1% and 72.3% of mutations in the training and blind datasets, respectively (Fig. 4; Table 1). The comparative performance across iterative non-redundant blind datasets suggested that the model was not overfitted.

Analysis of our model revealed that PncA-resistant mutations were associated with large changes in protein folding and stability (mCSM-Stability scores < -0.9 Kcal/mol; $p < 0.0001$, Welch Two Sample t-test) and conformational flexibility (DynaMut score < 0.78 Kcal/mol; $p < 0.0001$, Welch Two Sample t-test) or located in close proximity to the catalytic triad and substrate-binding site (<10.8 Å; $p < 0.0001$, Welch Two Sample t-test). Alternatively, susceptible mutations had a relative b-factor value of ≥ 3.19 ($p < 0.0001$, Welch Two Sample t-test), residue depth of ≥ 0.9 ($p < 0.0001$, Welch Two Sample t-test), distance from PZA greater than 11.9 Å and mild effects on protein stability (SDM scores ≥ 2.68 Kcal/mol; $p < 0.0001$, Welch Two Sample t-test).

Validation using Clinical Datasets. We next validated our model using variants reported in the recently published CRyPTIC dataset[24]. 355 *pncA* nsSNVs associated with PZA resistance were reported, of which 75 were not present in our training dataset. Our model correctly classified 79.2% of the mutations across the whole dataset

	Total nsSNVs	Resistant nsSNVs	correctly classified variants SUSPECT-PZA (%)	Susceptible nsSNVs	correctly classified variants SUSPECT-PZA (%)	PPV (%) (95% CI)	Accuracy (%)
Training dataset (70%)	426	213	159 (74.5)	213	182 (85.5)	83.7 (78.6–87.8)	80.1
Blind test dataset (30%)	184	92	56 (60.8)	92	77 (83.7)	78.9 (69.5–85.9)	72.3
CRyPTIC dataset[24]	355	325	266 (81.8)	30	15 (50.0)	94.7 (92.5–96.2)	79.2
CRyPTIC novel nsSNVs	75	67	67 (74.6)	8	4 (50.0)	92.6 (86.0–96.2)	72.0
Miotto *et al.* dataset[25]	98	92	82 (89.1)	6	2 (33.3)	95.4 (92.1–97.3)	84.8
Miotto novel nsSNVs	44	43	35 (81.4)	1	0	97.2 (96.8–97.6)	79.5
Stellenbosch University and CDC, USA nsSNVs[38]	8	5	3 (60.0)	3	3 (100)	100	75.0
Victorian TB novel nsSNVs	7	4	4 (100)	3	1 (33.3)	66. 7 (47.3–81.7)	71.4

Table 1. Evaluation metrics across the train and blind test datasets. Accuracy = (TP + TN)/(TP + TN + FP + FN); TP: True positives, TN: True Negatives, FP: False Positives, FN: False Negatives PPV: Positive predictive value, predicting PZA resistance (nsSNVs - non-synonymous single nucleotide variant).

(355 mutations), and 72.0% of those non-redundant in amino acid position with the training data (75 mutations). The positive predictive value was 94.7% (95% CI [92.5% to 96.2%]).

We also validated our empirical classifier using the dataset reported by Miotto *et al.*[25], which contained 98 nsSNVs graded by the confidence of their association with phenotypic drug resistance. 44 out of the 98 nsSNVs reported in the paper were not present in our training dataset. We accurately predicted the drug susceptibility of 84.8% of the polymorphism across the whole dataset (98 mutations), with an accuracy of 79.5% for those mutations not included in the training data (44 mutations). The positive predictive value was 95.4% (95% CI [92.1% to 97.3%]). We observed mutations such as Q10P (21 cases reported), W68G (16 cases reported) and I133T (17 cases reported) with 0.98 probability associated with resistant phenotype[22] and categorized as high confidence for association with resistance, moderate confidence for association with resistance and minimal confidence for association with resistance respectively[25] were all classified as resistant by our predictive model, highlighting the sensitivity of the prediction.

Mutations reported by Miotto *et al.*[25] under the "no association with resistance" category, including I31T, L35R and T47A were predicted as resistant, and I6L as susceptible. This is consistent with the available experimental data[24,28], highlighting the advantage, accuracy and versatility of our approach. A closer look into the different biophysical scores for the resistant associated mutations revealed that they had large predicted destabilizing values for protein conformational flexibility (I31T, −2.49 Kcal/mol) and stability (I31T, −3.46 Kcal/mol) and one was located very close to the catalytic triad (T47A, <6 Å).

Our predictive model was further validated on PZA DST screening at 100 μg/ml of clinical isolates from culture collections at Stellenbosch University, South Africa (865 isolates) and the Centers for Disease Control and Prevention (CDC), Atlanta, USA (185 isolates)[38]. They identified 49 isolates with a susceptible phenotype containing 8 nsSNVs. All nsSNVs with an MIC < 50 μg/ml were correctly classified by our model as susceptible (E37V, D110G, T114M). Whitfield and colleagues suggest that those isolates with an MIC > 50 μg/ml should be considered clinically resistant, of which our model classified three as resistant (A170V, V130A and L35R) and two as susceptible (V163A and V180I). Overall, our model had a 75% agreement with the DST results and a positive predictive value of 100%

Application within a Clinical Setting. In a prospective genomic sequencing and DST analysis of over 600 Victorian clinical TB isolates, 7 *pncA* variants were detected in 11 variants phenotypically resistant to PZA, none of which were present in our training dataset. Our model correctly classified five out of seven variants as resistant (71.4% accuracy). The remaining two mutations, G108V and Q10H, which were susceptible according to the DST results were predicted to confer resistance and consistent with other experimental findings[24,25,28]. Both variants, had a SNV frequency of <0.5, which is known to impact upon the reliability of the DST results. This highlights the potential clinical power of our model.

Expanding our analysis, four additional *pncA* mutations (S104R, V128G, Y95R and E15A) were identified in Victorian clinical TB isolates lacking DST results. Both S104R and V128G were predicted as resistant by our model, consistent with previously reported DST results[24–28]. The remaining two mutations, Y95R and E15A, have not been reported previously. Our model suggests both mutations to confer susceptibility to PZA.

SUSPECT-PZA webserver. We have developed a user-friendly, freely available web server SUSPECT-PZA (StrUctural Susceptibility PrEdiCTion on PZA), http://biosig.unimelb.edu.au/suspect_pza/, which is a database for all possible variants of PncA. There are two different input options (Fig. S2): the first one is the "Single Mutation" option which allows the users to input one mutation for analysis. The basic format required by the server for this input option is that the mutation must be specified as a text string containing the wild-type residue one-letter amino acid code, its corresponding position on the structure and the mutant one-letter amino acid code. The second option is the "Mutation List", which allows the user to upload a list of mutations, in the same specified format as above but in a file for batch processing (Fig. S3). Sample submission entries are available to assist users to submit their mutations for analysis and an additional help page via the top navigation bar.

Figure 5 shows a snapshot of the output page for the "Single Mutation" option. The web server displays the prediction outcome (Resistant / Susceptible) along with details of the user input data, information on the wildtype residue environment and features used for prediction. In addition, there is an interactive 3D viewer, built using NGL[39], which allows analysis of non-covalent inter-residue interactions for the position specified in the input calculated using Arpeggio[30] for both wild-type and mutant structures. The results for the "Mutation List" option is summarized in a downloadable table. The users can access details of individual mutation as shown in Fig. S4. There is a 3D viewer at the bottom of the page in which the residues in the input list is colored according to the predicted effect (Fig. S5).

Discussion

PZA was discovered in 1948 in an *in vivo* screen of nicotinamide derivatives in a structure-activity relationship study[40] and used as anti-tuberculosis drug in 1952 for the first time. Till the 1970's PZA was used as a second-line drug to treat TB, until they discovered the sterilizing activity and reduction in treatment duration in combination with isoniazid and rifampicin. There has been a lot of studies conducted since then and with the continued usage of the drug to treat TB, there has been an increased incidence of resistance associated with it. Being an important first-line drug, accurate and rapid evaluation of PZA susceptibility is crucial for successful management of patients with either susceptible or drug-resistant TB. The existing molecular phenotypic tests are considered poorly reliable, expensive, and has a long turnaround time. To account for this situation there is an urgent requirement to develop a rapid, reliable and affordable molecular PZA DST. As resistance mutations are spread all over the length of the PncA protein, it is quite challenging to develop a new method. In this study, we establish a novel computational methodology to better understand the structural and functional consequences of drug resistance mutations by exploiting the protein's 3D structure. Using supervised machine learning algorithm, we developed an empirical tool to determine novel drug resistance in PncA followed by a database which has information on all possible variants of PncA.

The primary focus of our work is on missense non-synonymous mutations as these typically have more subtle molecular effects that can be harder to predict, than in-frame and frameshift indel mutations that have a much larger deleterious effect on PncA structure and function and are all classed as high-confidence resistant mutations. The structure-based tools implement the concept of graph-based signatures to predict the effect on single point mutations for protein stability. To assess changes in conformational flexibility, graph-based signatures were integrated with normal mode analysis to predict the impact on the protein structure. Scores for these features which were calculated as change in Gibb's free energy ($\Delta\Delta G$) provided important molecular information on resistant mutations, signifying larger effects on protein folding and dynamics and minimal effect on PZA binding affinity. Interpreting the results, we observed, resistance mutations were seen to affect protein activity and function through destabilization of the protein structure and conformation. It even helped in correlating earlier findings where resistant isolates were not associated with a loss of bacterial fitness[41] due to the fact that PncA was involved in nicotinamide recycling pathway rather than in its synthesis. These structural insights have been used to guide clinical decisions for novel PZA mutations[23].

Phenotypic DST which is the current "gold standard", which encompasses methods like Wayne and Bactec MGIT 960, suffers from poor reproducibility. Discrepancies among the results lead to considerable doubt over the clinical significance of the method. Next-generation sequencing based diagnostics can be an alternative for innovative tools to reduce false detection of PZA resistance cases and fast and accurate detection of drug resistance by molecular DST[42]. In the past couple of years researchers have used different techniques to come up with a better and consistent methodology to detect and determine resistance in PZA. Stoffels *et al.*[41] conducted an elaborate study on 14-year complete capture of clinical isolates, where he found frequency of spontaneous acquired resistance to be 10^{-5} bacilli *in vitro*. Miotto *et al.* 2014 work generated the minimum dataset of mutations that should be included in any molecular test for PZA, paving the way for predicting PZA resistance using new genome-based technologies[22]. This was followed by Farhat *et al.* 2016 comprehensive web-based dataset[43]. Though all these approaches were a step up from the existing phenotypic DST, they do not provide information on novel variants. The advantage with our database is it provides information on all possible variants for PncA. This data provides a basis for use as part of any molecular DST, needed for the valid interpretation of data generated by massive sequencing approaches.

Interestingly, comparing performance of SUSPECT-PZA across datasets used to train earlier methods, we observed that the weakest performance was across variants classified as susceptible. However, many of these mutations have been observed in clinically resistant isolates. Our biophysical analysis and SUSPECT-PZA predictions would be consistent with these mutations potentially being misclassified previously.

We also compared our empirical models output to the "revised DST" of Miotto *et al.*[22], where they accounted for enzymatic activity and structural analysis to adjust for possible errors in phenotypic DST. There were 178 missense mutations listed, of which 162 were labelled resistant (R) and 17 were labelled susceptible (S). Our model predicted 88.9% (144/162) of the resistant mutations and 58.8% (10/17) of the susceptible mutations accurately. The positive predictive value was 95.4% (95% CI [92.1% to 97.3%]). The primary divergence from the Miotto classifications was in predicting susceptible mutations. This is likely due to discrepancies in phenotypic and molecular DST results from different laboratory setups[16]. For example, mutations reported as susceptible in the "revised DST" like L159V, F81S, A102V, T135S, T168I and A46V were unanimously reported as resistant in other studies[24,26–28]. Our predictive tool also predicts them to be resistant and hence, proves to be more reliable, reproducible, free to use and a fast alternative to the existing gold standard methods.

This study highlights the power of using computational prediction of the structural consequences of variants in PncA to identify likely pyrazinamide resistance mutations, a critically important first-line drug in the treatment of tuberculosis. This approach, however, is not limited to pncA and has been developed for application to other antimicrobial agents like bedaquiline[44], a last line resort to treat multi-drug and extremely drug resistant

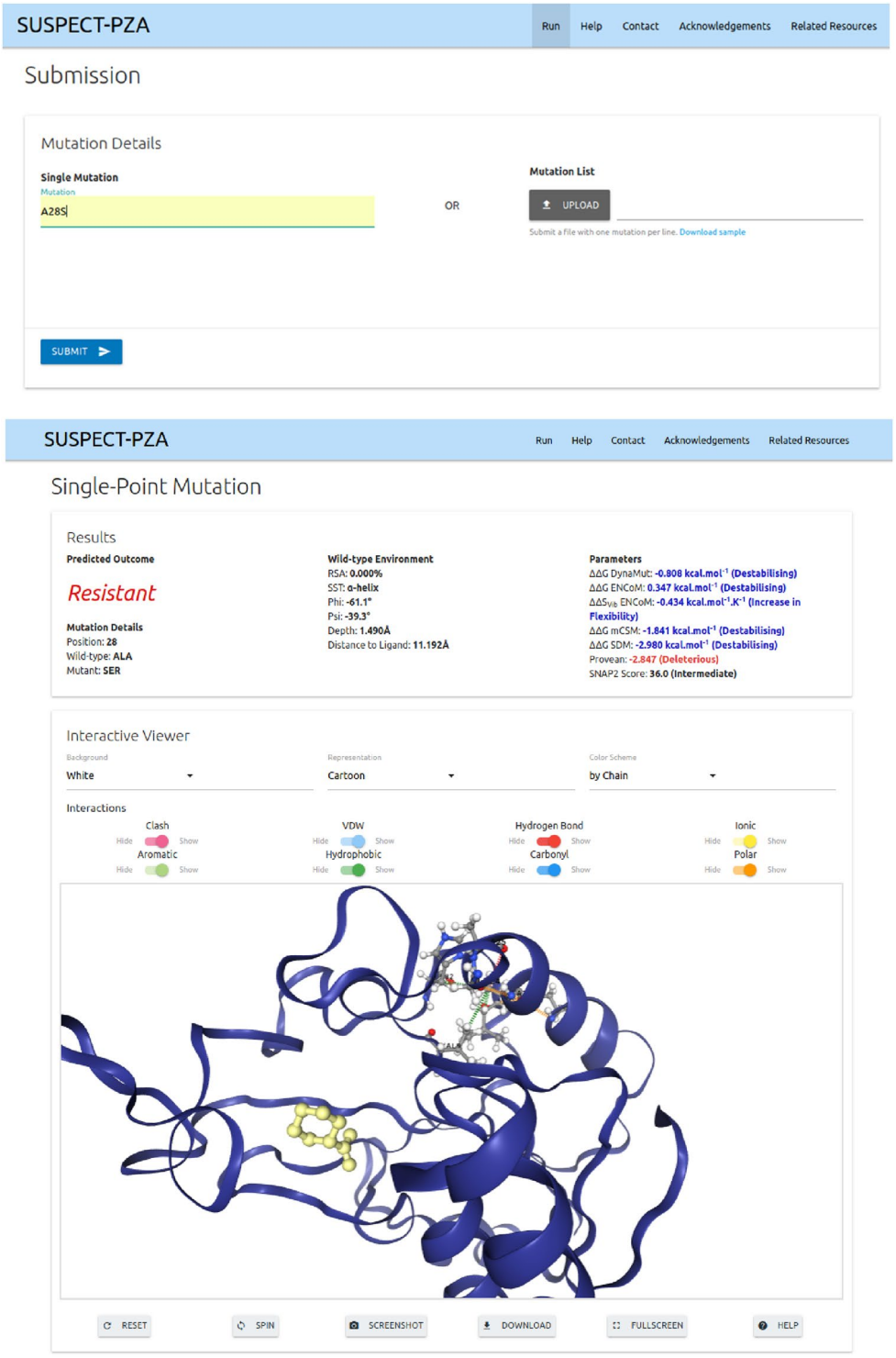


Figure 5. SUSPECT-PZA webserver Single point mutation prediction result page. The predicted outcome of the submitted mutation is displayed along with complimentary information on features used to aid in the development of the tool. The interactive 3D viewer allows user to further analyze non-covalent interactions for both wild type and mutant residues on the protein. A variety of controllers are provided to customize molecule representation.

TB. A major advantage of our tool is that it was built using a very well-balanced dataset. In case of mutations reported as both susceptible and resistant in the same or different datasets, we looked for frequency of occurrence and clinical information. We have extensively evaluated the method through both cross-validation and independent non-redundant blind tests, which provide a measure of a methods applicability and robustness. Across all test sets the method performed equally well, providing strong confidence in the approach. As with all machine

learning approaches, the availability of more phenotypic and clinical data will enable the development and validation of stronger approaches. This will be an iterative approach moving forward. The other aspect to improving our predictive model is through the inclusion of new features or parameters. We have shown previously that this approach can even capture strain dependent variations in resistant patterns[23]. While we did not have the data available to build into our current model, we next aim to integrate lineage specific information, which will enable more refined and personalized predictions. This comprehensive web server can be used in clinical settings as an improved diagnostic tool to help realize the power of whole genome sequencing diagnostic approaches.

Methods

Data set. A list of 610 nonsynonymous single-nucleotide mutations (nsSNVs) of pncA was obtained from the GMTV (Genome-wide Mycobacterium tuberculosis Variation) Database Project[26], Tuberculosis Drug Resistance Mutation Database[27], and saturation mutagenesis[28]. The clinical validation datasets used in the paper were from CRyPTIC[24] and Miotto *et al.*[25].

Modelling the biophysical consequences of missense mutations. We have developed a comprehensive *in silico* mutational analysis platform that uses graph-based signatures to represent the 3D structure of a protein and quantitatively predict the molecular consequences of point mutations on protein structure, function and interactions[30,33–36,45]. This has been used to characterize and preemptively identify likely resistance mutations in drug targets[23,37,46–54]. Using these tools, we assessed the molecular consequences of each mutation on the structure of PncA and drug activation.

The experimental crystal structure of holo-wild-type PncA (PDB ID: 3PL1)[29] was minimized in Prime, and PZA docked into the active site using Glide (Schrödinger Suite). The effects of mutations on PncA folding and stability were assessed using SDM[55], mCSM-Stability[33] and DUET[34], and their effects on protein flexibility and conformational was predicted using normal mode analysis by DynaMut[35]. The effect of the changes on the binding affinity of PZA towards PncA were predicted using mCSM-Lig[36,56]. These approaches are novel machine-learning algorithms. We also included structural information of the wild-type residue, including relative solvent accessibility, residue depth, secondary structure and dihedral angles of the PncA chain φ (phi) and ψ (psi). Additionally, SNAP2[31] and PROVEAN[32] were used to provide additional evolutionary information. Moreover, the scores calculated for the various structural and sequence-based features are independent of pH and temperature.

Machine learning. Here we used the Random Forest binary classifier using the Weka toolkit[57] to train our predictive models. Random Forest is an ensemble-learning robust classification algorithm, in which multiple decision trees are included over a random subset of features and decide the output via majority voting. The model was trained using 10-fold cross-validation and performance evaluated by area under the Receiver Operating Characteristic (AUROC) curve, precision and accuracy. Further validation of the models was performed using a blind-test set of 184 mutations, which were non-redundant at the position-level with mutations in the training set. Analysis of the final model revealed a set of structural features that distinguished between susceptible and resistant pncA point mutations.

Webserver development. The server front-end was built using materialize CSS framework version 1.0.0, while the backend was built in Python via the Flask framework (version 0.12.2). It is hosted on a Linux server running Apache.

Sequencing and DST of clinical isolates. Genomic DNA was extracted according to the mechanical cell disruption and ethanol precipitation method outlined in Votintseva 2015[58] with slight modifications. Briefly, no pre-treatment was used and approximately $3 \times 1\,\mu L$ loops of culture were dispersed in 700 µL TE buffer (Sigma Aldrich) as the starting material. The precipitated DNA pellet was only washed once and resuspended into 50 µL EB Buffer (Qiagen) at 55 °C for 10 minutes with regular vortexing. Finally, samples were centrifuged 3 min at 13,000 rpm and 45 µL of DNA extract was transferred into a clean tube for downstream processing. Each extract was interrogated for *Mycobacterium tuberculosis* viability by inoculating 15 µL of DNA extract into MGIT tube (Becton Dickinson, UK) and incubated in the Bactec MGIT 960 system (Becton Dickinson, UK). Unique dual indexed libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina). Libraries were sequenced on the Illumina NextSeq. 500 with 150-cycle paired end chemistry as described by the manufacturer's protocols.

Sequences were aligned to H37Rv (NC_0009623.3) and small nucleotide variations (SNV) mutations in *pncA* were identified using LoFreq (http://csb5.github.io/lofreq/). SNVs with a frequency > 0.6 were used to compare the genotype of isolates to the phenotype observed using standard laboratory methods for PZA susceptibility testing.

Received: 11 July 2019; Accepted: 28 November 2019;
Published online: 05 February 2020

References

1. WHO. Global Tuberculosis Report, Executive Summary, 2018. https://www.who.int/tb/publications/global_report/tb18_ExecSum_web_4Oct18.pdf?ua=1 (2018).
2. Heifets, L. & Lindholm-Levy, P. Pyrazinamide sterilizing activity *in vitro* against semidormant Mycobacterium tuberculosis bacterial populations. *The American review of respiratory disease* **145**, 1223–1225, https://doi.org/10.1164/ajrccm/145.5.1223 (1992).
3. Tarshis, M. S. & Weed, W. A. Jr. Lack of significant *in vitro* sensitivity of Mycobacterium tuberculosis to pyrazinamide on three different solid media. *American review of tuberculosis* **67**, 391–395 (1953).
4. Dawson, R. *et al.* Efficiency and safety of the combination of moxifloxacin, pretomanid (PA-824), and pyrazinamide during the first 8 weeks of antituberculosis treatment: a phase 2b, open-label, partly randomised trial in patients with drug-susceptible or drug-resistant pulmonary tuberculosis. *Lancet (London, England)* **385**, 1738–1747, https://doi.org/10.1016/s0140-6736(14)62002-x (2015).

5. Veziris, N. *et al*. A once-weekly R207910-containing regimen exceeds activity of the standard daily regimen in murine tuberculosis. *American journal of respiratory and critical care medicine* **179**, 75–79, https://doi.org/10.1164/rccm.200711-1736OC (2009).
6. Juma, S. P. *et al*. Underestimated pyrazinamide resistance may compromise outcomes of pyrazinamide containing regimens for treatment of drug susceptible and multi-drug-resistant tuberculosis in Tanzania. *BMC infectious diseases* **19**, 129, https://doi.org/10.1186/s12879-019-3757-1 (2019).
7. Chang, K. C., Yew, W. W. & Zhang, Y. Pyrazinamide susceptibility testing in Mycobacterium tuberculosis: a systematic review with meta-analyses. *Antimicrobial agents and chemotherapy* **55**, 4499–4505, https://doi.org/10.1128/aac.00630-11 (2011).
8. Chedore, P., Bertucci, L., Wolfe, J., Sharma, M. & Jamieson, F. Potential for erroneous results indicating resistance when using the Bactec MGIT 960 system for testing susceptibility of Mycobacterium tuberculosis to pyrazinamide. *Journal of clinical microbiology* **48**, 300–301, https://doi.org/10.1128/jcm.01775-09 (2010).
9. Hewlett, D. Jr., Horn, D. L. & Alfalla, C. Drug-resistant tuberculosis: inconsistent results of pyrazinamide susceptibility testing. *Jama* **273**, 916–917 (1995).
10. Hoffner, S. *et al*. Proficiency of drug susceptibility testing of Mycobacterium tuberculosis against pyrazinamide: the Swedish experience. *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease* **17**, 1486–1490, https://doi.org/10.5588/ijtld.13.0195 (2013).
11. Miller, M. A., Thibert, L., Desjardins, F., Siddiqi, S. H. & Dascal, A. Testing of susceptibility of Mycobacterium tuberculosis to pyrazinamide: comparison of Bactec method with pyrazinamidase assay. *Journal of clinical microbiology* **33**, 2468–2470 (1995).
12. Pandey, S., Newton, S., Upton, A., Roberts, S. & Drinkovic, D. Characterisation of pncA mutations in clinical Mycobacterium tuberculosis isolates in New Zealand. *Pathology* **41**, 582–584 (2009).
13. Simons, S. O. *et al*. Validation of pncA gene sequencing in combination with the mycobacterial growth indicator tube method to test susceptibility of Mycobacterium tuberculosis to pyrazinamide. *Journal of clinical microbiology* **50**, 428–434, https://doi.org/10.1128/jcm.05435-11 (2012).
14. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nature medicine* **2**, 662–667 (1996).
15. Konno, K., Feldmann, F. M. & McDermott, W. Pyrazinamide susceptibility and amidase activity of tubercle bacilli. *The American review of respiratory disease* **95**, 461–469, https://doi.org/10.1164/arrd.1967.95.3.461 (1967).
16. Zhang, Y., Wade, M. M., Scorpio, A., Zhang, H. & Sun, Z. Mode of action of pyrazinamide: disruption of Mycobacterium tuberculosis membrane transport and energetics by pyrazinoic acid. *The Journal of antimicrobial chemotherapy* **52**, 790–795, https://doi.org/10.1093/jac/dkg446 (2003).
17. Shi, W. *et al*. Pyrazinamide inhibits trans-translation in Mycobacterium tuberculosis. *Science (New York, N.Y.)* **333**, 1630–1632, https://doi.org/10.1126/science.1208813 (2011).
18. Shi, W. *et al*. Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis. *Emerging microbes & infections* **3**, e58, https://doi.org/10.1038/emi.2014.61 (2014).
19. Yee, M., Gopal, P. & Dick, T. Missense Mutations in the Unfoldase ClpC1 of the Caseinolytic Protease Complex Are Associated with Pyrazinamide Resistance in Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy* **61**, https://doi.org/10.1128/aac.02342-16 (2017).
20. Zhang, Y., Zhang, J., Cui, P., Zhang, Y. & Zhang, W. Identification of Novel Efflux Proteins Rv0191, Rv3756c, Rv3008, and Rv1667c Involved in Pyrazinamide Resistance in Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, **61**, https://doi.org/10.1128/aac.00940-17 (2017).
21. Hirano, K., Takahashi, M., Kazumi, Y., Fukasawa, Y. & Abe, C. Mutation in pncA is a major mechanism of pyrazinamide resistance in Mycobacterium tuberculosis. *Tubercle and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease* **78**, 117–122 (1997).
22. Miotto, P. *et al*. Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *mBio* **5**, e01819–01814, https://doi.org/10.1128/mBio.01819-14 (2014).
23. Karmakar, M. *et al*. Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy. *American journal of respiratory and critical care medicine* **198**, 541–544, https://doi.org/10.1164/rccm.201712-2572LE (2018).
24. Allix-Beguec, C. *et al*. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *The New England journal of medicine* **379**, 1403–1415, https://doi.org/10.1056/NEJMoa1800474 (2018).
25. Miotto, P. *et al*. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *The European respiratory journal*, **50**, https://doi.org/10.1183/13993003.01354-2017 (2017).
26. Chernyaeva, E. N. *et al*. Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. *BMC genomics* **15**, 308, https://doi.org/10.1186/1471-2164-15-308 (2014).
27. Sandgren, A. *et al*. Tuberculosis drug resistance mutation database. *PLoS medicine* **6**, e2, https://doi.org/10.1371/journal.pmed.1000002 (2009).
28. Yadon, A. N. *et al*. A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nature communications* **8**, 588, https://doi.org/10.1038/s41467-017-00721-2 (2017).
29. Petrella, S. *et al*. Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PloS one* **6**, e15785, https://doi.org/10.1371/journal.pone.0015785 (2011).
30. Jubb, H. C. *et al*. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of molecular biology* **429**, 365–371, https://doi.org/10.1016/j.jmb.2016.12.004 (2017).
31. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC genomics* **16**(Suppl 8), S1, https://doi.org/10.1186/1471-2164-16-s8-s1 (2015).
32. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PloS one* **7**, e46688, https://doi.org/10.1371/journal.pone.0046688 (2012).
33. Pires, D. E., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics (Oxford, England)* **30**, 335–342, https://doi.org/10.1093/bioinformatics/btt691 (2014).
34. Pires, D. E., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research* **42**, W314–319, https://doi.org/10.1093/nar/gku411 (2014).
35. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic acids research* **46**, W350–w355, https://doi.org/10.1093/nar/gky300 (2018).
36. Pires, D. E., Blundell, T. L. & Ascher, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific reports* **6**, 29575, https://doi.org/10.1038/srep29575 (2016).
37. Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G. & Furnham, N. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Scientific reports* **8**, 15356, https://doi.org/10.1038/s41598-018-33370-6 (2018).
38. Whitfield, M. G. *et al*. Mycobacterium tuberculosis pncA Polymorphisms That Do Not Confer Pyrazinamide Resistance at a Breakpoint Concentration of 100 Micrograms per Milliliter in MGIT. *Journal of clinical microbiology* **53**, 3633–3635, https://doi.org/10.1128/jcm.01001-15 (2015).
39. Rose, A. S. *et al*. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics (Oxford, England)* **34**, 3755–3758, https://doi.org/10.1093/bioinformatics/bty419 (2018).
40. Kushner, S. *et al*. Experimental chemotherapy of tuberculosis; substituted nicotinamides. *The Journal of organic chemistry* **13**, 834–836, https://doi.org/10.1021/jo01164a008 (1948).

41. Stoffels, K., Mathys, V., Fauville-Dufaux, M., Wintjens, R. & Bifani, P. Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy* **56**, 5186–5193, https://doi.org/10.1128/aac.05385-11 (2012).
42. Koser, C. U. *et al*. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS pathogens* **8**, e1002824, https://doi.org/10.1371/journal.ppat.1002824 (2012).
43. Farhat, M. R. *et al*. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *American journal of respiratory and critical care medicine* **194**, 621–630, https://doi.org/10.1164/rccm.201510-2091OC (2016).
44. Karmakar, M. *et al*. Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PloS one* **14**, e0217169, https://doi.org/10.1371/journal.pone.0217169 (2019).
45. Pires, D. E., Chen, J., Blundell, T. L. & Ascher, D. B. *In silico* functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Scientific reports* **6**, 19848, https://doi.org/10.1038/srep19848 (2016).
46. Ascher, D. B. *et al*. Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for. *RNA*. *Scientific reports* **4**, 4765, https://doi.org/10.1038/srep04765 (2014).
47. Kano, F. S. *et al*. The Presence, Persistence and Functional Properties of Plasmodium vivax Duffy Binding Protein II Antibodies Are Influenced by HLA Class II Allelic Variants. *PLoS Negl. Trop. Dis.* **10**, e0005177, https://doi.org/10.1371/journal.pntd.0005177 (2016).
48. Phelan, J. *et al*. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14**, 31, https://doi.org/10.1186/s12916-016-0575-9 (2016).
49. Silvino, A. C. *et al*. Variation in Human Cytochrome P-450 Drug-Metabolism Genes: A Gateway to the Understanding of Plasmodium vivax Relapses. *PloS one* **11**, e0160172, https://doi.org/10.1371/journal.pone.0160172 (2016).
50. Albanaz, A. T. S., Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin. Drug Discov* **12**, 553–563, https://doi.org/10.1080/17460441.2017.1322579 (2017).
51. Park, Y. *et al*. Essential but Not Vulnerable: Indazole Sulfonamides Targeting Inosine Monophosphate Dehydrogenase as Potential Leads against Mycobacterium tuberculosis. *ACS infectious diseases* **3**, 18–33, https://doi.org/10.1021/acsinfecdis.6b00103 (2017).
52. Singh, V. *et al*. The Inosine Monophosphate Dehydrogenase, GuaB2, Is a Vulnerable New Bactericidal Drug Target for Tuberculosis. *ACS infectious diseases* **3**, 5–17, https://doi.org/10.1021/acsinfecdis.6b00102 (2017).
53. Hawkey, J. *et al*. Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. *Microbial. Genomics* **4**, -, https://doi.org/10.1099/mgen.0.000165 (2018).
54. Holt, K. E. *et al*. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* **50**, 849–856, https://doi.org/10.1038/s41588-018-0117-9 (2018).
55. Worth, C. L., Preissner, R. & Blundell, T. L. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research* **39**, W215–222, https://doi.org/10.1093/nar/gkr363 (2011).
56. Pires, D. E. & Ascher, D. B. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic acids research* **44**, W557–561, https://doi.org/10.1093/nar/gkw390 (2016).
57. Hall, M. *et al*. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18, https://doi.org/10.1145/1656274.1656278 (2009).
58. Votintseva, A. A. *et al*. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *Journal of clinical microbiology* **53**, 1137–1143, https://doi.org/10.1128/jcm.03073-14 (2015).

Acknowledgements

M.K. and C.M.H.R. were funded by the Melbourne Research Scholarship. Funding for genomic sequencing was provided by the Department of Health and Human Services, Victoria. D.B.A. was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). This work was supported in part by the Victorian Government's OIS Program.

Author contributions

M.K. performed the analysis and along with C.H.M.R. developed the analysis tool. K.H. and J.D. contributed to data collected and analysis. D.B.A. conceived, designed and supervised the project. All authors contributed to manuscript writing and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58635-x.

Correspondence and requests for materials should be addressed to D.B.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix N

A comprehensive computational platform to guide drug development using graph-based signature methods

Chapter 7

A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods

Douglas E. V. Pires, Stephanie Portelli, Pâmela M. Rezende, Wandré N. P. Veloso, Joicymara S. Xavier, Malancha Karmakar, Yoochan Myung, João P. V. Linhares, Carlos H. M. Rodrigues, Michael Silk, and David B. Ascher

Abstract

High-throughput computational techniques have become invaluable tools to help increase the overall success, process efficiency, and associated costs of drug development. By designing ligands tailored to specific protein structures in a disease of interest, an understanding of molecular interactions and ways to optimize them can be achieved prior to chemical synthesis. This understanding can help direct crucial chemical and biological experiments by maximizing available resources on higher quality leads. Moreover, predicting molecular binding affinity within specific biological contexts, as well as ligand pharmacokinetics and toxicities, can aid in filtering out redundant leads early on within the process. We describe a set of computational tools which can aid in drug discovery at different stages, from hit identification (EasyVS) to lead optimization and candidate selection (CSM-lig, mCSM-lig, Arpeggio, pkCSM). Incorporating these tools along the drug development process can help ensure that candidate leads are chemically and biologically feasible to become successful and tractable drugs.

Key words Graph-based signatures, mCSM, Mutation, Protein-ligand, Interatomic interactions, Docking, Drug development

1 Introduction

Structure-guided drug development uses knowledge of the three-dimensional structure of the biological target to more efficiently guide the design of small molecule binders. While it has become an integral strategy for both lead generation and optimization, the application of computational tools to take advantage of the explosion in structural information has often required specialist knowledge and resources and in some cases has been limited to commercial software.

Zoltán Gáspári (ed.), *Structural Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, vol. 2112,
https://doi.org/10.1007/978-1-0716-0270-6_7,

Using the concept of graph-based signatures, we have developed a robust, user-friendly, and freely accessible platform to analyze protein structures and interactions [1–12] and guide disease characterization [13–28] and drug development [29–32]. These include methods to perform virtual screening (EasyVS), score protein-small molecule docking solutions (CSM-lig [3]), look at all the molecular interactions being made (Arpeggio [7]), identify mutations that are likely to affect compound binding (mCSM-lig [5]), and characterize the pharmacokinetic and toxicity properties of the proposed molecules (pkCSM [33, 34]). These have been successfully employed in a number of drug development projects [30–32, 35–37] and together comprise a powerful platform that allows users to enhance their structure-guided drug development efforts (Fig. 1). Here we discuss how this platform can be leveraged to guide drug development.

2 Materials

Here we present four structure-based tools to help guide drug development. For each method, users are required to provide:

1. **Wild-type protein structure in PDB format**: For all methods, a wild-type structure in the Protein Data Bank [38] format must be provided to perform the analysis. This can be an experimentally solved structure previously deposited into the Protein Data Bank (www.rcsb.org or http://www.ebi.ac.uk/pdbe/) or a model, for instance, obtained by comparative homology modeling. We have previously shown that homology models built using templates down to 25% sequence identity do not significantly affect the accuracy of the methods [9, 10]. For Arpeggio, CSM-lig, and mCSM-lig, the protein structure file needs to include the ligand of interest, either already present in the experimental structure or computationally docked into the binding site. PDB structures are required to have a valid chain identifier (*see* **Note 1**), a single conformation (multiple occupancies need to be filtered out; *see* **Note 2**), and a single model, in case of NMR structures (*see* **Note 3**).
2. **Three-letter code of the ligand of interest**: When a structure of a protein-ligand complex is provided to the predictive web servers (CSM-lig and mCSM-lig), users will be asked to provide a three-letter code that identifies the residue ID for that ligand within the PDB file, according to the PDB format standards. In addition to the three-letter code, CSM-lig also requires the canonical SMILES of the compound of interest for additional property calculations. Several tools are available to aid users to convert between small molecule formats. These include stand-alone packages such as OpenBabel [39] and Avogadro [40].

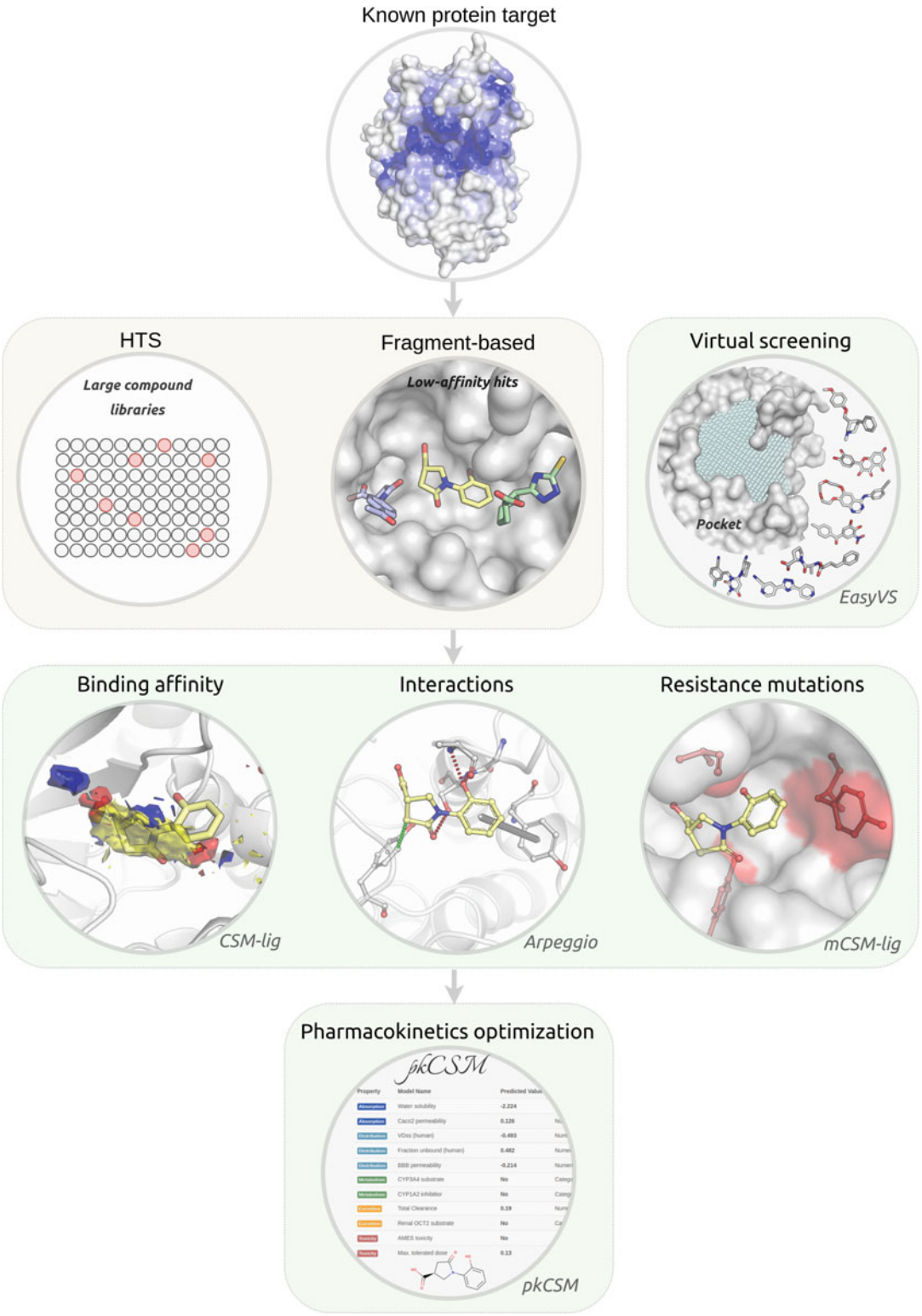


Fig. 1 A structure-based computational platform to guide drug development. To complement and support traditional experimental approaches, including high-throughput screening (HTS) and fragment-based drug discovery, this in silico platform supports hit identification via virtual screening, methods to better understand protein–small molecule interactions, affinity and effects of mutations, as well as the optimization of pharmacokinetic properties

3 Methods

3.1 Performing Automated Docking with EasyVS

1. Virtual screening is a powerful, high-throughput technique for computationally screening large libraries of small molecules (often in the order of millions) in order to identify those ligands which are most likely to bind to a drug target protein. When compared to traditional screening methods, this leads to significantly higher hit rates that can proceed to lead optimization [41, 42]. It can, however, be computationally intensive and usually requires specialist knowledge. EasyVS provides an easy-to-use web interface at http://biosig.unimelb.edu.au/easyvs/, allowing users to rapidly set up and analyze their virtual screening results.
2. Users can upload the structure of the protein target of interest as either a PDB file or by providing the PDB ID of a previously solved experimental structure. Any ligands, ions, or water molecules already bound to the provided structure will be disregarded.
3. On the following step, the provided PDB file or identifier will be processed, and pockets will be automatically detected using Ghecom [43] (Fig. 2a-1). Users can either select one of the identified pockets to determine the docking grid (the three-dimensional space where the ligands will be docked into) or provide specific grid coordinates and size (Fig. 2a-2).
4. Users then need to select the ligand library they want to screen, which includes libraries of purchasable compounds, natural products, or FDA-approved drugs (Fig. 2b). These can be further filtered based upon their molecular properties (e.g., Lipinski's rule of five [44] or the rule of three) or grouped by similarity.
5. The selected molecules will then be docked into the selected docking grid (Fig. 2c-1), and the top 20 poses per ligand can be downloaded. The server also provides an interactive visualization tool to compare ligand docking poses (Fig. 2c-2). The example on this figure shows the docking poses for ligands docked to the Ribosome-Inactivating Protein Ricin A (PDB ID: 1BR5). While poses are sorted by predicted affinity (kcal/mol) using autodock's scoring function, users can evaluate docking poses with alternative approaches, such as CSM-lig [3].

3.2 Predicting Protein-Small Molecule Affinity with CSM-lig

1. Following virtual screening or docking, the affinity of the top docked ligand poses can be quantified using CSM-lig. This is a machine learning-based tool which acts as a scoring function and enables the numerical affinity comparison between poses. It is implemented via an easy-to-use web interface at http://biosig.unimelb.edu.au/csm_lig, which is compatible with most operating systems and browsers.

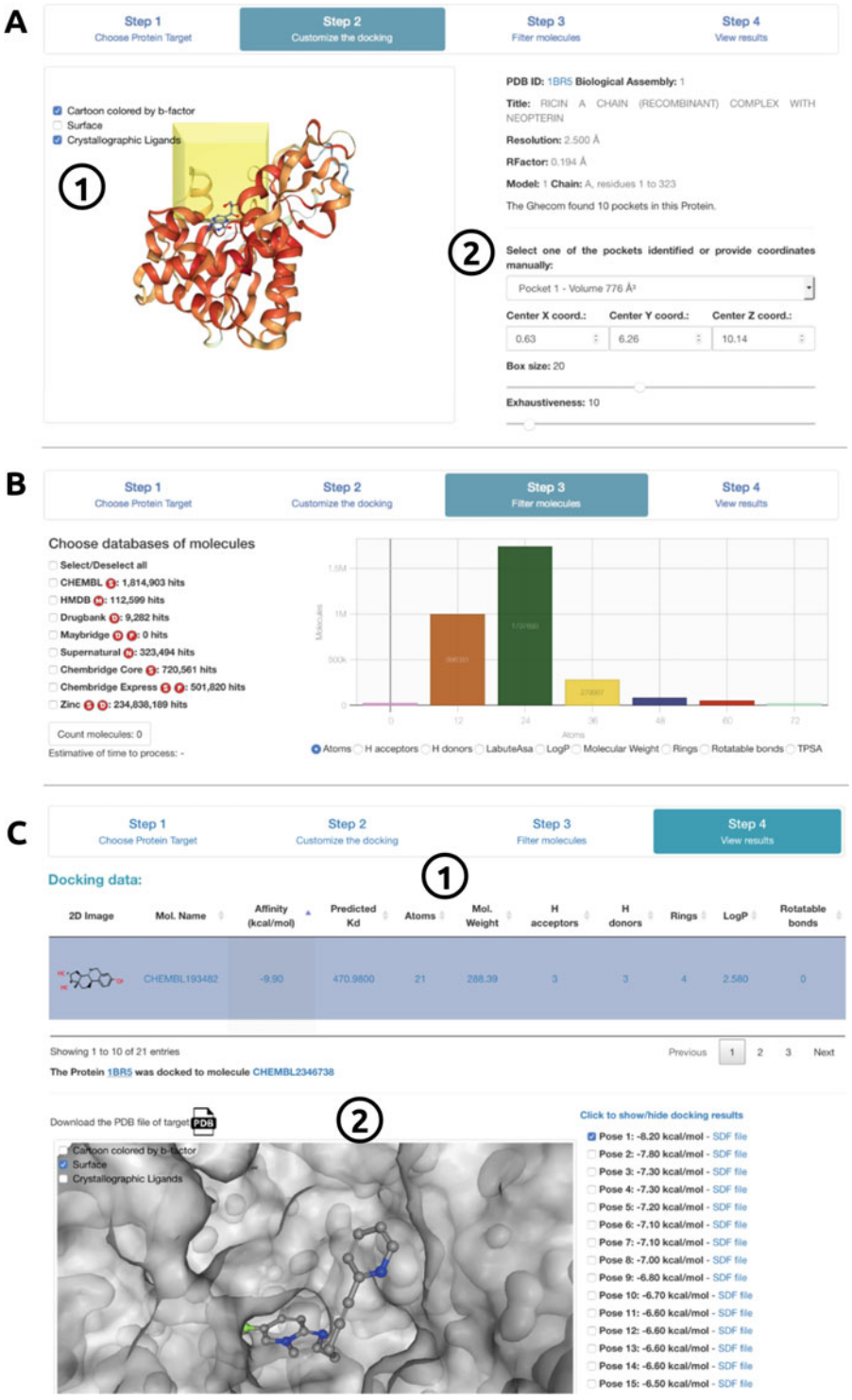


Fig. 2 Automated docking with EasyVS. After choosing a target of interest, EasyVS will automatically identify pockets (**a-1**) and allow user to further customize the docking protocol (**a-2**). A range of ligand libraries can be selected for docking (**b**), including FDA-approved drugs, purchasable compounds, and natural products, which can be further filtered based on physicochemical properties. Docking results are shown in tabular format (**c-1**), depicting ligands, their properties, and docking scores. An interactive viewer allows users to inspect the best poses for each ligand (**c-2**)

2. By selecting the "Predict" tab, users are presented with two job options, "Single Structure" and "Multiple Structures."
3. For "Single Structure" prediction, provide (Fig. 3a-1) the protein-small molecule complex you would like to evaluate the pose of in PDB format (Fig. 3a-2), the three-letter code for the small molecule (as in the provided PDB file) and (Fig. 3a-3) and the SMILES string of the small molecule.
4. Alternatively, for "Multiple Structures," provide two files. The first file (Fig. 3a-4) is a compressed zip file with all protein-small molecule PDB files you would like to evaluate. These could be, for instance, different poses or conformations for a given protein-ligand complex or multiple different complexes. The second (Fig. 3a-5) is a tab-separated file with the following information for each uploaded complex in the .zip file: (a) structure file name (file in PDB format), (b) three-letter code for the small molecule (as in the structure file), and (c) canonical SMILES for the small molecule.
5. The output prediction page for the "Single Structure" jobs depicted in Fig. 1b presents (Fig. 3b-1) the predicted affinity (as $-\log_{10}(\textit{affinity})$ in molar, meaning a compound with an affinity predicted as 1 nM would have a predicted value of 9). The example presented in the figure and the web server shows the affinity prediction for the ligand Zanamivir bound to human sialidase-2 (PDB ID: 2F0Z). For this complex, CSM-lig generates a score of 12.6, denoting very high affinity (larger numbers denote higher affinity). A depiction figure of the small molecule is shown, together with calculated properties, including molecular weight (in Da) and partition coefficient (log *P*), among others (Fig. 3b-2). An interactive visualization of the protein-small molecule complex is also exhibited (Fig. 3b-3). The interatomic non-covalent interactions between protein and small molecule are also calculated and are available as a downloadable Pymol [45] session (Fig. 3b-4). Pharmacokinetics and toxicity predictions by pkCSM for the provided small molecule are also available by clicking on the red button at the bottom-left corner of the results page.
6. The output for "Multiple Structures" jobs are shown in tabular format (Fig. 3c-1), depicting predicted affinity values, SMILES identifying the molecules and their calculated molecular properties. These results are available as a tabular file and can be downloaded (Fig. 3c-2).

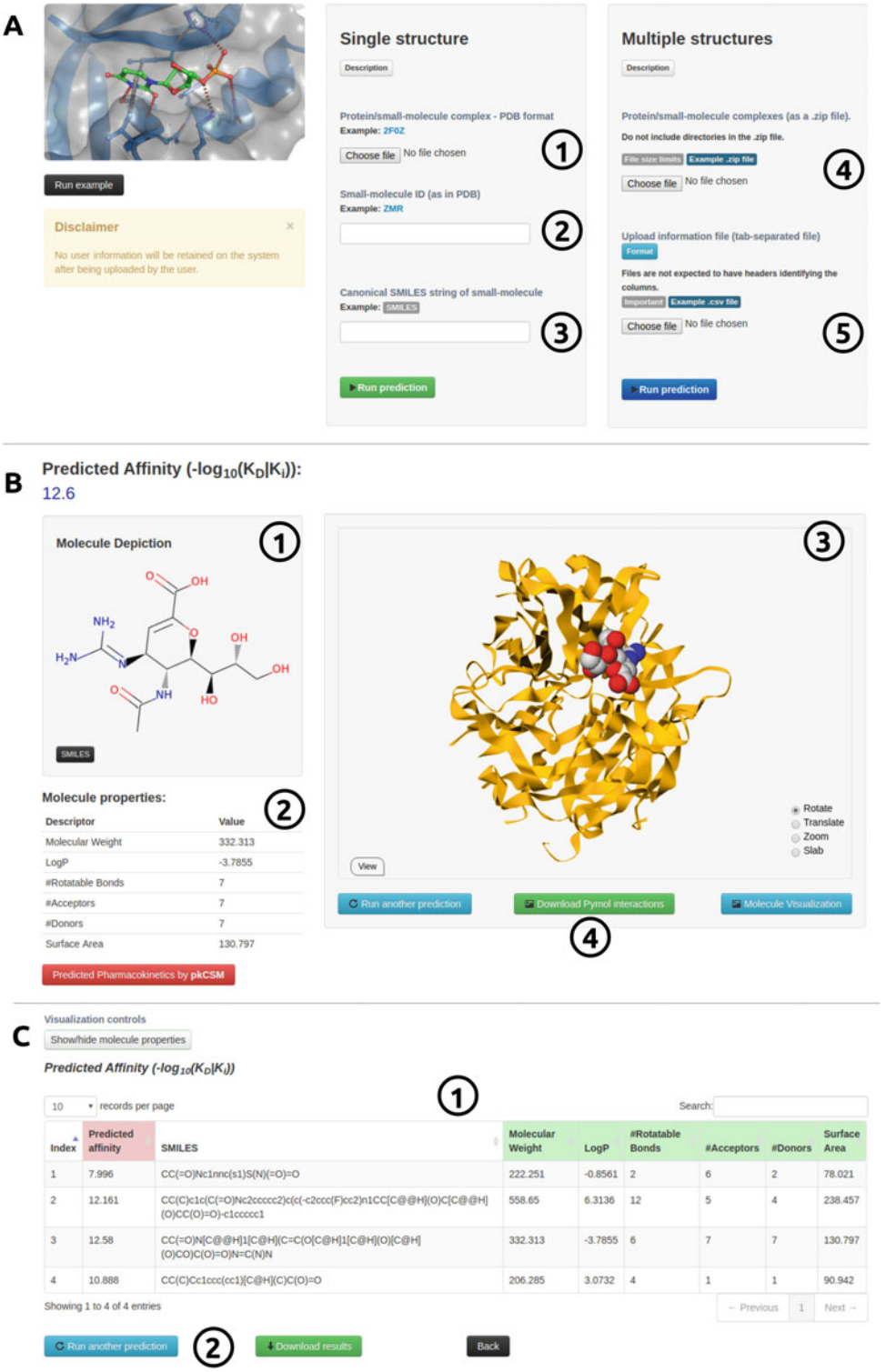


Fig. 3 CSM-lig submission and results web interface. The submission page (**a**) allows users to provide either single or multiple protein-ligand complexes for evaluation. The results page for single complex/pose assessment (**b**) provides the calculated affinity, ligand properties and depiction, as well as an interactive visualization of the complex. For multiple poses, CSM-lig provides the predicted affinities in a downloadable tabular format, together with ligand properties (**c**)

3.3 Depicting and Analyzing Protein-Small Molecule Interactions with Arpeggio

1. Once a structure of the target protein with the candidate molecule is available, either through experimental determination or docking or other alternative approach (for instance, those combining blind docking with molecular dynamics like the Wrap 'n' Shake method [46]), Arpeggio enables the visualization of intermolecular interactions occurring between the lead and its target. During lead optimization, Arpeggio can therefore be used to understand the mechanism of binding and guide medicinal chemistry efforts.
2. Arpeggio is freely available as a user-friendly web interface and is compatible with multiple operating systems and browsers. Open up the prediction server, http://biosig.unimelb.edu.au/arpeggioweb/, on a web browser of your preference.
3. Provide the complexed protein structure of interest by either uploading it as a PDB file or providing the PDB ID of the experimentally solved structure in complex with the ligand of interest (Fig. 4a-1).
4. Select the ligand or ligands of interest under the "Heteroatom" selection heading to calculate all molecular interactions being made by that ligand (Fig. 4b-1; *see* **Note 4**).
5. The results page will show an interactive image of all the molecular interactions made by the ligand(s) selected (Fig. 5a) and a table with a count of the total number of specific molecular interactions being made, including hydrophobic interactions, hydrogen bonds, pi-interactions, and ionic interactions (Fig. 4c).
6. A Pymol session file (PSE file) containing the submitted PDB file and all of the calculated interactions can be downloaded and opened in Pymol to enable visualization of the interaction network in 3D and to facilitate high-quality image generation for manuscripts (Fig. 5b).

3.4 Predicting the Effects of Mutations on Small Molecule Affinity with mCSM-lig

1. During lead optimization, it is important to consider how genetic diversity might affect the binding of candidate molecules and, in particular, if resistance is likely to arise. mCSM-lig uses graph-based signatures to calculate the change upon mutation in small molecule binding affinity. In order to run a prediction, open up the mCSM-lig server at http://biosig.unimelb.edu.au/mcsm_lig/ on a web browser of your preference (the web server is compatible with the most common operating systems and browsers).
2. Users are required to provide the protein structure in complex with the ligand of interest by either uploading a PDB file or supplying a valid four-letter code PDB accession code of a deposited experimental structure (Fig. 6a-1). Users also need to provide the mutation information, the mutation chain, the

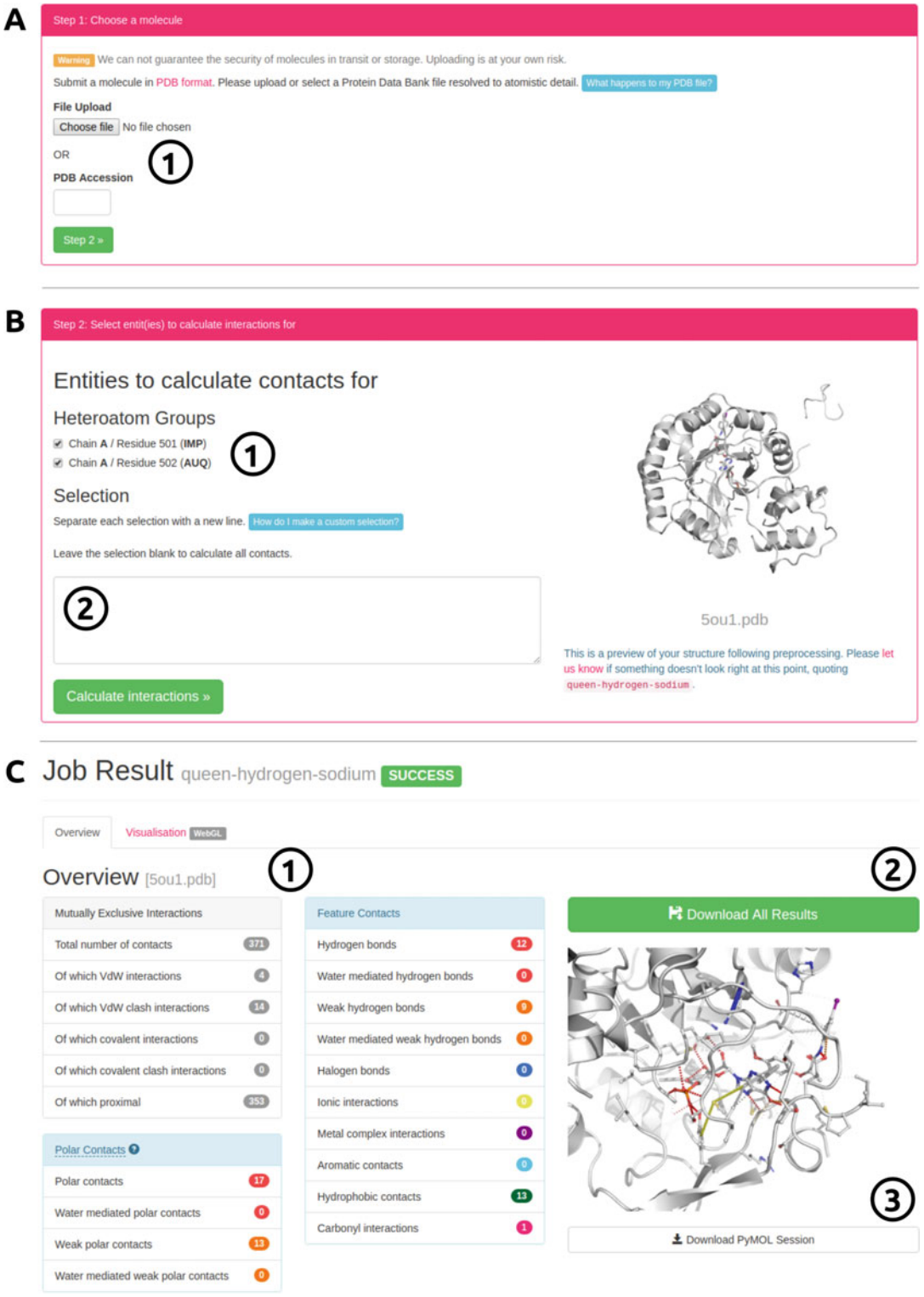


Fig. 4 Arpeggio submission and results web interface. (**a**) The submission page allows users to either provide their own PDB file or an accession code of a deposited experimental structure of the protein of interest. By selecting the molecule of interest (**b**), all molecular interactions will be calculated and displayed (**c**)

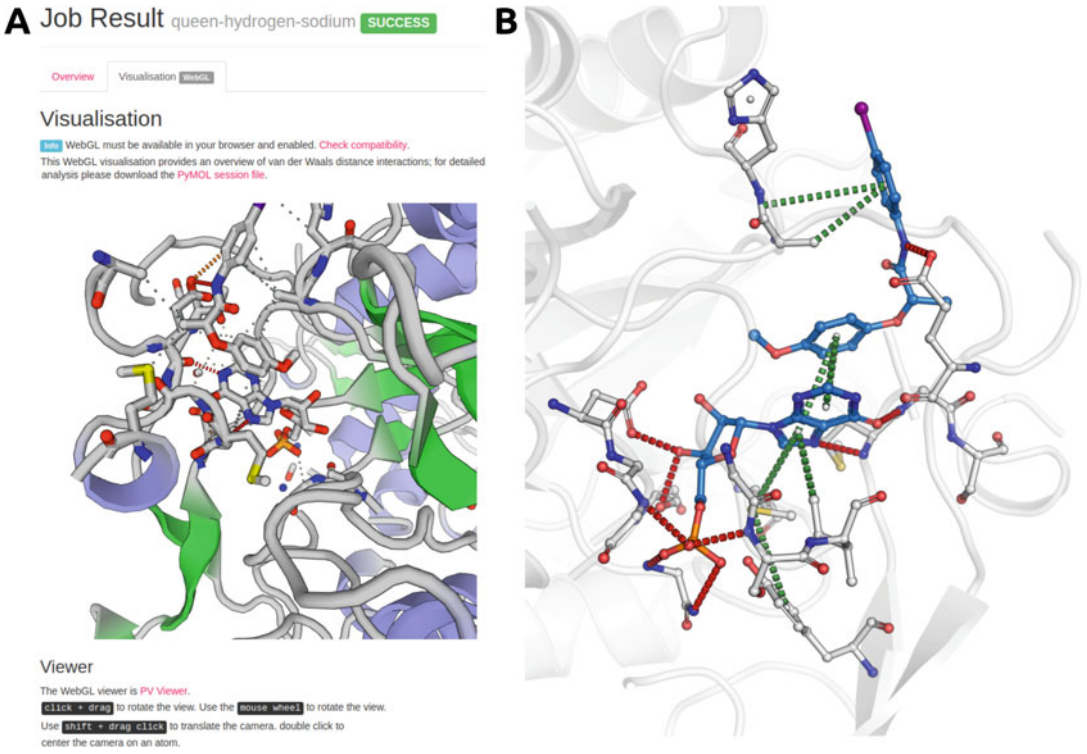


Fig. 5 Molecular interaction visualization using Arpeggio. The molecular interactions calculated by Arpeggio can be visualized either online (**a**) or by downloading the PSE file for visualization in Pymol (**b**)

three-letter code of the ligand of interest in the PDB file, and the approximate binding affinity (in nM) (Fig. 6a-2). If the binding affinity is not available, this can be approximated using CSM-lig. The mCSM-lig values do not vary significantly across most biologically relevant binding affinities.

3. After processing, the results page is shown (Fig. 6b-1), which includes information about the mutation and the predicted effects on the ligand binding affinity. An interactive molecular visualization is shown, allowing users to inspect the wild-type residue environment (Fig. 6b-2).
4. Predicted effects are outputted as the log fold change in binding affinity, in which negative values denote destabilizing mutations and positive values, stabilizing ones. The example shown in Fig. 6 and the web server depicts the prediction for a mutation on the HIV-1 protease bound to an inhibitor. Mutation from Aspartic Acid to Asparagine on residue position 30 is predicted to considerably reduce protein-ligand affinity. While users should interpret the values in the context of the protein system being studied, for competitive binding inhibitors, it is often important to consider the relative effect of a mutation on not only inhibitor binding but also the competitive ligand. This

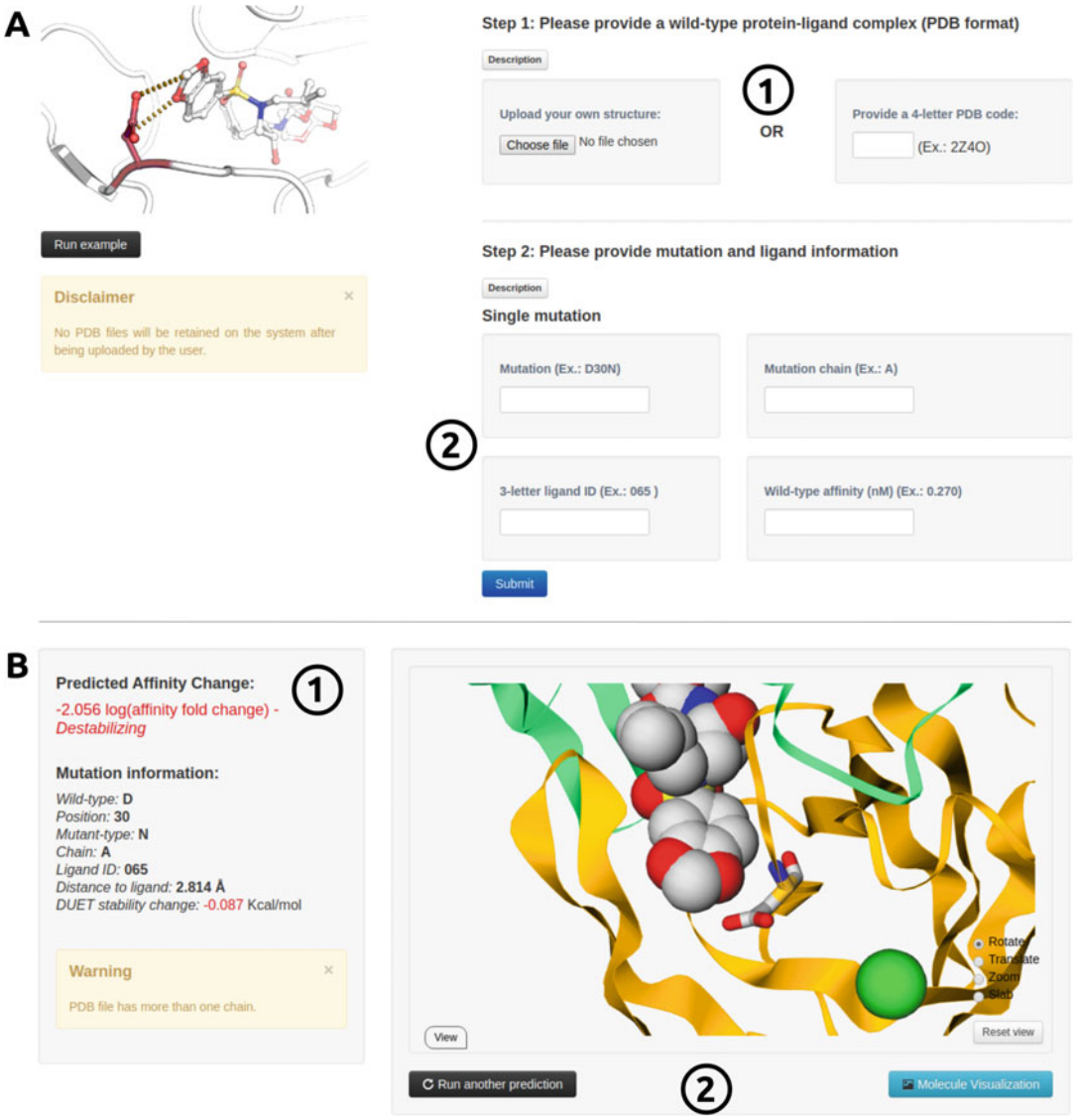


Fig. 6 mCSM-lig submission and results web interface. To predict the effects of a mutation on protein-ligand affinity, users need to provide a protein-ligand structure of interest (**a-1**) as well as mutation and ligand information (**a-2**). Once the calculations have finished, the results page will show the predicted change in ligand binding affinity (**b-1**) as well as an interactive visualization of the mutated residue within its molecular environment (**b-2**)

can be done by submitting a structure of the protein containing the ligand. Resistance mutations are more likely to affect, or have a larger effect, on inhibitor binding affinity than the natural ligand. This has been used to successfully preemptively guide detection of likely resistance variants [29–31, 47–53].

4 Notes

1. The chain ID for the provided PDB file is a mandatory field for CSM-Lig and mCSM-Lig, and blank characters are not allowed. It is possible that homology modeling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. There are several tools available to perform this task.
2. Another source of error comes from multiple occupancies, common in high-resolution experimental X-ray crystal structures. Multiple occupancies should first be filtered out, with the highest occupancy conformation normally selected.
3. NMR experimental structures often contain multiple models. It is an important practice to filter NMR structures, selecting a single model. The predictive tool will show a warning message in case multiple models are identified.
4. Arpeggio will sometimes fail if the PDB file contains an element with upper and lower case letters (e.g., Fe as opposed to FE). These can be altered using a text editor.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; the Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P., P.M.R.]; and the Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.].

References

1. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
2. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42 (Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
3. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44 (W1):W557–W561. https://doi.org/10.1093/nar/gkw390
4. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based

signatures. Nucleic Acids Res 44(W1): W469–W473. https://doi.org/10.1093/nar/gkw458

5. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
6. Pires DE, Chen J, Blundell TL, Ascher DB (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
7. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
8. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45(W1): W229–W235. https://doi.org/10.1093/nar/gkx439
9. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46(W1): W127–W132. https://doi.org/10.1093/nar/gky375
10. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300
11. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
12. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45 (W1):W241–W246. https://doi.org/10.1093/nar/gkx236
13. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, Rattenberry E, Simpson MA, Trembath RC, Weber A, Woodward ER, Donaldson A, Blundell TL, Latif F, Maher ER (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096
14. Jubb H, Blundell TL, Ascher DB (2015) Flexibility and small pockets at protein-protein interfaces: new insights into druggability. Prog Biophys Mol Biol 119(1):2–9. https://doi.org/10.1016/j.pbiomolbio.2015.01.009
15. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
16. Coelho MB, Ascher DB, Gooding C, Lang E, Maude H, Turner D, Llorian M, Pires DE, Attig J, Smith CW (2016) Functional interactions between polypyrimidine tract binding protein and PRI peptide ligand containing proteins. Biochem Soc Trans 44(4):1058–1065. https://doi.org/10.1042/BST20160080
17. Kano FS, Souza-Silva FA, Torres LM, Lima BA, Sousa TN, Alves JR, Rocha RS, Fontes CJ, Sanchez BA, Adams JH, Brito CF, Pires DE, Ascher DB, Sell AM, Carvalho LH (2016) The presence, persistence and functional properties of Plasmodium vivax Duffy binding protein II antibodies are influenced by HLA class II allelic variants. PLoS Negl Trop Dis 10(12):e0005177. https://doi.org/10.1371/journal.pntd.0005177
18. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenco CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
19. Silvino AC, Costa GL, Araujo FC, Ascher DB, Pires DE, Fontes CJ, Carvalho LH, Brito CF, Sousa TN (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172
20. White RR, Ponsford AH, Weekes MP, Rodrigues RB, Ascher DB, Mol M, Selkirk ME, Gygi SP, Sanderson CM, Artavanis-Tsakonas K (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11):

e1005977. https://doi.org/10.1371/journal.ppat.1005977

21. Casey RT, Ascher DB, Rattenberry E, Izatt L, Andrews KA, Simpson HL, Challis B, Park SM, Bulusu VR, Lalloo F, Pires DEV, West H, Clark GR, Smith PS, Whitworth J, Papathomas TG, Taniere P, Savisaar R, Hurst LD, Woodward ER, Maher ER (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
22. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montano B, Blundell TL, Ascher DB (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
23. Ramdzan YM, Trubetskov MM, Ormsby AR, Newcombe EA, Sui X, Tobin MJ, Bongiovanni MN, Gras SL, Dewson G, Miller JML, Finkbeiner S, Moily NS, Niclis J, Parish CL, Purcell AW, Baker MJ, Wilce JA, Waris S, Stojanovski D, Bocking T, Ang CS, Ascher DB, Reid GE, Hatters DM (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
24. Soardi FC, Machado-Silva A, Linhares ND, Zheng G, Qu Q, Pena HB, Martins TMM, Vieira HGS, Pereira NB, Melo-Minardi RC, Gomes CC, Gomez RS, Gomes DA, Pires DEV, Ascher DB, Yu H, Pena SDJ (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2(1):7. https://doi.org/10.1038/s41525-017-0009-4
25. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27(10):1715–1729. https://doi.org/10.1101/gr.226589.117
26. Trezza A, Bernini A, Langella A, Ascher DB, Pires DEV, Sodi A, Passerini I, Pelo E, Rizzo S, Niccolai N, Spiga O (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58(12):5320–5328. https://doi.org/10.1167/iovs.17-22158
27. Andrews KA, Ascher DB, Pires DEV, Barnes DR, Vialard L, Casey RT, Bradshaw N, Adlard J, Aylwin S, Brennan P, Brewer C, Cole T, Cook JA, Davidson R, Donaldson A, Fryer A, Greenhalgh L, Hodgson SV, Irving R, Lalloo F, McConachie M, McConnell VPM, Morrison PJ, Murday V, Park SM, Simpson HL, Snape K, Stewart S, Tomkins SE, Wallis Y, Izatt L, Goudie D, Lindsay RS, Perry CG, Woodward ER, Antoniou AC, Maher ER (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55(6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
28. Hnizda A, Fabry M, Moriyama T, Pachl P, Kugler M, Brinsa V, Ascher DB, Carroll WL, Novak P, Zaliova M, Trka J, Rezacova P, Yang JJ, Veverka V (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32(6):1393–1403. https://doi.org/10.1038/s41375-018-0073-5
29. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
30. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, Arora K, Ioerger TR, Sacchettini J, Rizzi M, Donini S, Blundell TL, Ascher DB, Rhee K, Breda A, Zhou N, Dartois V, Jonnala SR, Via LE, Mizrahi V, Epemolu O, Stojanovski L, Simeons F, Osuna-Cabello M, Ellis L, MacKenzie CJ, Smith AR, Davis SH, Murugesan D, Buchanan KI, Turner PA, Huggett M, Zuccotto F, Rebollo-Lopez MJ, Lafuente-Monasterio MJ, Sanz O, Diaz GS, Lelievre J, Ballell L, Selenski C, Axtman M, Ghidelli-Disse S, Pflaumer H, Bosche M, Drewes G, Freiberg GM, Kurnick MD, Srikumaran M, Kempf DJ, Green SR, Ray PC, Read K, Wyatt P, Barry CE 3rd, Boshoff HI (2017) Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against Mycobacterium tuberculosis. ACS Infect Dis 3(1):18–33. https://doi.org/10.1021/acsinfecdis.6b00103
31. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, Keri G, Ascher DB, Mondesert G, Vocat A, Lupien A, Sommer R, Vermet H, Lagrange S, Buechler J, Warner DF, McKinney JD, Pato J, Cole ST, Blundell TL, Rizzi M, Mizrahi V (2017) The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis 3(1):5–17. https://doi.org/10.1021/acsinfecdis.6b00102

32. Trapero A, Pacitto A, Singh V, Sabbah M, Coyne AG, Mizrahi V, Blundell TL, Ascher DB, Abell C (2018) Fragment-based approach to targeting inosine-5′-monophosphate dehydrogenase (IMPDH) from Mycobacterium tuberculosis. J Med Chem 61(7):2806–2822. https://doi.org/10.1021/acs.jmedchem.7b01622
33. Pires DE, Blundell TL, Ascher DB (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 58 (9):4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104
34. Pires DEV, Kaminskas LM, Ascher DB (2018) Prediction and optimization of pharmacokinetic and toxicity properties of the ligand. Methods Mol Biol 1762:271–284. https://doi.org/10.1007/978-1-4939-7756-7_14
35. Sigurdardottir AG, Winter A, Sobkowicz A, Fragai M, Chirgadze D, Ascher DB, Blundell TL, Gherardi E (2015) Exploring the chemical space of the lysine-binding pocket of the first kringle domain of hepatocyte growth factor/scatter factor (HGF/SF) yields a new class of inhibitors of HGF/SF-MET binding. Chem Sci 6(11):6147–6157. https://doi.org/10.1039/c5sc02155c
36. Ascher DB, Jubb HC, Pires DE, Ochi T, Higueruelo A, Blundell TL (2015) Protein-protein interactions: structures and druggability. In: Scapin G, Patel D, Arnold E (eds) Multifaceted roles of crystallography in modern drug discovery. NATO science for peace and security series a: chemistry and biology. Springer, Netherlands, pp 141–163. https://doi.org/10.1007/978-94-017-9719-1_12
37. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
39. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33
40. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 4(1):17. https://doi.org/10.1186/1758-2946-4-17
41. Ascher DB, Crespi GA, Ng HL, Morton CJ, Parker MW (2008) Novel therapeutic approaches to treat Alzheimer's disease and memory disorders. J Proteomics Bioinform 1:464–476
42. Chai SY, Yeatman HR, Parker MW, Ascher DB, Thompson PE, Mulvey HT, Albiston AL (2008) Development of cognitive enhancers based on inhibition of insulin-regulated aminopeptidase. BMC Neurosci 9(Suppl 2):S14. https://doi.org/10.1186/1471-2202-9-S2-S14
43. Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 78(5):1195–1211. https://doi.org/10.1002/prot.22639
44. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46 (1–3):3–26
45. Schrodinger, LLC (2015) The PyMOL molecular graphics system, version 1.8
46. Balint M, Jeszenoi N, Horvath I, van der Spoel D, Hetenyi C (2017) Systematic exploration of multiple drug binding sites. J Cheminform 9(1):65. https://doi.org/10.1186/s13321-017-0255-6
47. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
48. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, Coeck N, Hill-Cawthorne GA, Nair MB, Mallard K, Ramsay A, Campino S, Hibberd ML, Pain A, Rigouts L, Clark TG (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14(1):31. https://doi.org/10.1186/s12916-016-0575-9
49. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4. https://doi.org/10.1099/mgen.0.000165
50. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage

and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50 (6):849–856. https://doi.org/10.1038/s41588-018-0117-9

51. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, Denholm JT, Ascher DB (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 198 (4):541–544. https://doi.org/10.1164/rccm.201712-2572LE
52. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8 (1):15356. https://doi.org/10.1038/s41598-018-33370-6
53. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, Arumugam S, Rajan L, Ebenezer M, Ascher DB, Arnold E, Blundell TL (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1

Appendix O

Identifying genotype-phenotype correlations via integrative mutation analysis

Chapter 1

Identifying Genotype–Phenotype Correlations via Integrative Mutation Analysis

Edward Airey, Stephanie Portelli, Joicymara S. Xavier, Yoo Chan Myung, Michael Silk, Malancha Karmakar, João P. L. Velloso, Carlos H. M. Rodrigues, Hardik H. Parate, Anjali Garg, Raghad Al-Jarf, Lucy Barr, Juliana A. Geraldo, Pâmela M. Rezende, Douglas E. V. Pires, and David B. Ascher

Abstract

Mutations in protein-coding regions can lead to large biological changes and are associated with genetic conditions, including cancers and Mendelian diseases, as well as drug resistance. Although whole genome and exome sequencing help to elucidate potential genotype–phenotype correlations, there is a large gap between the identification of new variants and deciphering their molecular consequences. A comprehensive understanding of these mechanistic consequences is crucial to better understand and treat diseases in a more personalized and effective way. This is particularly relevant considering estimates that over 80% of mutations associated with a disease are incorrectly assumed to be causative. A thorough analysis of potential effects of mutations is required to correctly identify the molecular mechanisms of disease and enable the distinction between disease-causing and non–disease-causing variation within a gene. Here we present an overview of our integrative mutation analysis platform, which focuses on refining the current genotype–phenotype correlation methods by using the wealth of protein structural information.

Key words Genotype–phenotype correlations, Graph-based signatures, mCSM, Mutation, Protein structure, Protein interactions

1 Introduction

Proteins are versatile molecules, responsible for orchestrating a wide range of biological processes. They comprise a single polypeptide chain of amino acids, which folds in 3D space into dynamic structures. How a protein folds is important for determining its functions, including activities and interactions with other molecules. These structures are highly coordinated and conserved across evolution, and small perturbations in the amino acid sequence can disrupt these shapes, functions, and interactions [1, 2]. While

Hugh Cartwright (ed.), *Artificial Neural Networks*, Methods in Molecular Biology, vol. 2190,
https://doi.org/10.1007/978-1-0716-0826-5_1,

missense mutations, causing a change to a single amino acid, are generally less structurally disruptive than nonsense mutations, their effects are highly variable and can be wide-ranging, making their molecular consequences harder to determine. Despite their subtle effects, missense substitutions are related with many different genetic conditions, including cancer, Mendelian diseases, and the emergence of drug resistance.

The introduction of a missense mutation can have many molecular effects, including altering how the protein folds, its dynamics, posttranslational modifications, half-life, localization, activity, and molecular interactions [3]. When analyzing a new mutation, an integrative approach is therefore important to consider the effects it might have on all of these aspects. This enables the identification of specific functional, and structural changes imparted by the mutations, which is essential for a molecular understanding. It can also explain why mutations in the same protein might lead to different diseases, why mutations might cluster in 3D space and how those genetic changes present phenotypically.

Although many assume that an unfavorable phenotype (e.g., pathogenic, drug-resistant) is the result of large, overall destabilizing mutations, mutations with milder effects are often more prevalent in a population, as they are generally under less selective pressure [4, 5]. For example, by assessing mutations in three different tuberculosis proteins that lead to resistance, we have shown that the most frequent resistant mutations were more likely to be associated with overall mild functional effects, and associated reduced fitness cost, allowing for increased prevalence within the bacterial population [4].

Experimentally elucidating the biophysical effects of mutations is an expensive and time-consuming task, usually limited to a few variants in proteins with amenable assays. Over the years, the accumulation of information of experimentally characterized mutations has enabled the development and improvement of computational mutational analysis tools [6]. These computational platforms have shown to be invaluable assets to decipher genotype–phenotype correlations in cancer [7–19], Mendelian diseases [20–26], and detection of antimicrobial resistance [4, 15, 27–35], guiding clinical decisions and driving further research. Here, we introduce a general computational pipeline that uses in silico biophysical predictions and machine learning approaches to harness the wealth of available biological and protein structural information and give insights into genotype–phenotype correlation for clinical use [10].

The mutation cutoff scanning matrix (mCSM) platform is the only comprehensive collection of in silico tools for quantitatively predicting the effects of missense mutations on protein folding, structure, dynamics, and interactions. It includes tools which calculate all possible molecular interactions (Arpeggio [36]), account for changes in protein stability (mCSM-Stability [37], SDM [38],

DUET [39], mCSM-membrane [40], dynamics (DynaMut [41]), protein interactions with other proteins (mCSM-PPI [37], mCSM-PPI2 [42], mCSM-AB [43], mCSM-AB2 [44], mmCSM-AB [45], nucleic acids (mCSM-DNA [37], mCSM-NA [46]), and small molecule ligands (mCSM-lig [47], CSM-lig [48]).

These tools were built using the concept of graph-based signatures [49, 50], which represent the geometry and physicochemical properties of the wild-type protein structure environment as a network or graph, composed of a series of nodes, describing the local mutation environment, and edges, describing the distances between interacting "layers" of surrounding residues. Information on the mutation is captured using the pharmacophore change between the wild-type and the mutant residue, including whether charges or hydrogen donors/acceptors have been gained or lost [37].

This platform allows for accurate biophysical predictions, which, when complemented with other protein analytical tools, can provide a detailed landscape on the specific mutational effects on a protein. We have implemented these within an analytical and supervised machine learning predictive pipeline (Fig. 1), to enable easy and fast characterization of novel mutations and their likely

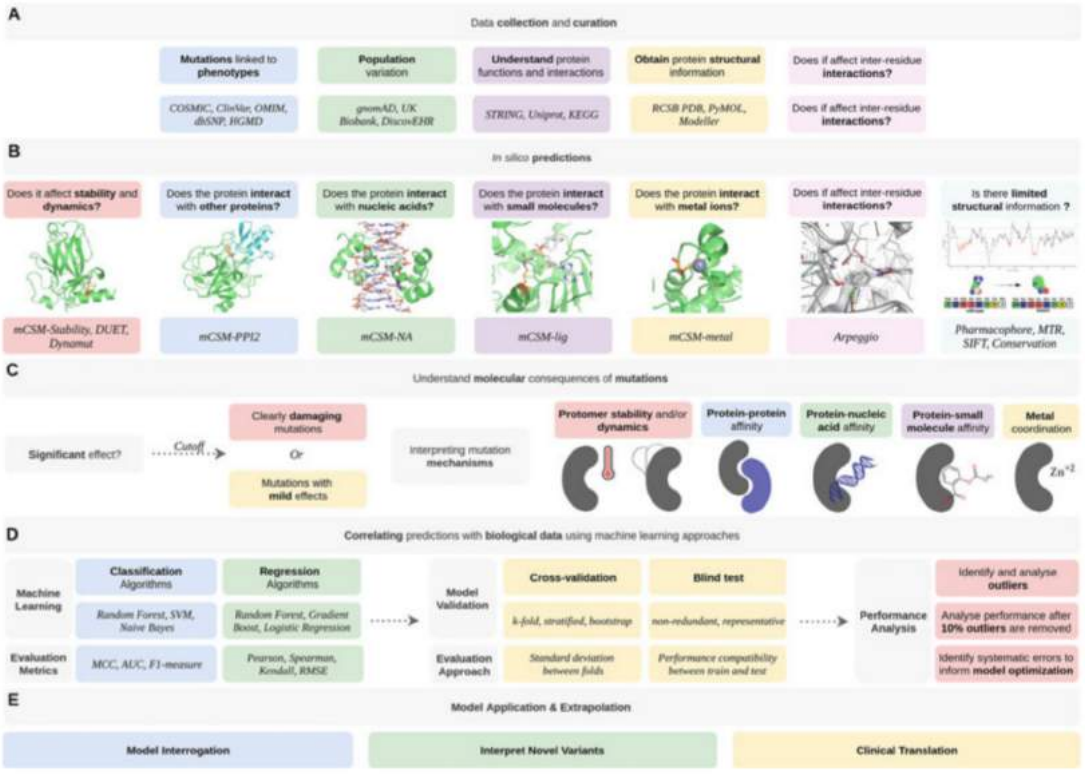


Fig. 1 An overview of the mechanistic characterization of mutations and their biological consequences, to guide the development of tools to predict phenotypic outcomes

clinical phenotypes. This approach has been shown to have big implications in diagnostic and personalized medicine in the post-genomic era.

2 Materials

2.1 Data Curation

2.1.1 Mutation Curation

The foremost requirement for training a machine learning model is appropriate high-quality experimental/clinical data, with suitable representation of the classes under comparison. For human disease, a wealth of freely accessible collections of curated data exist. Previously reported mutations through publications and functional studies are available from dbSNP [51], the largest freely available repository of genetic variation. Variants with evidence of pathogenicity can be viewed from the Human Gene Mutation Database (HGMD) [52] and ClinVar [53], and from disease-specific datasets such as the Catalogue of Somatic Mutations in Cancer (COSMIC). Standing variation is available from genomic sequencing efforts of healthy populations, including over 140,000 healthy humans in gnomAD [54] and 50,000 whole exomes currently available in UK Biobank [55].

When combining data from multiple sources, it is important that all datapoints are comparable. If using genetic coordinates, they should be found on the same assembly of the genome (e.g., GRCh38 vs GRCh37). The mutations themselves (whether reported as genetic or amino acid changes) must be reported on the same transcript, as most genes have multiple reported coding sequences.

2.1.2 Protein Structure Curation

The sequence and functional information for a specific protein can be obtained from Uniprot (https://www.uniprot.org/) [56]. To run the mCSM tools we need crystallographic structures, which can be downloaded from the Protein Data Bank (PDB;http://www.rcsb.org/) [57] or generated via homology modeling or molecular docking (to run mCSM-PPI, mCSM-Lig, or mCSM-NA). Once we have the variant information collected from the resources in Subheading 2.1.1, we map these variants on to the identified protein structures to help visualize the spread and identify potential hotspots, which is easily done using visualization software such as PyMol, as it enables selection of residues being mutated in a 3D manner.

2.2 An Overview of Computational Tools to Analyze Missense Mutations

Over the past two decades there has been an unprecedented growth in both computational power and the amount of biological data available. This has facilitated the development of numerous sequence (Table 1) and structural (Table 2) based computational tools to guide mutation characterization.

Table 1
Available sequence-based predictive tools for mutation analysis

Protein stability and dynamics	
Method	**Corr.[a]**
I-Mutant 2.0	0.62
Auto-Mute	0.64[a]
MUpro	0.75
DynaMine	0.63
DDGun	0.49
INPS-MD/3D	0.58
iStable	0.56[b]
iPTREEE - STAB	0.70
ProMaya	0.79

[a]Pearson's correlation
[b]MCC

Table 2
Available structure-based predictive tools for mutation analysis

Protein stability and dynamics		Protein–protein affinity		Protein–nucleic acid affinity		Protein–small molecule affinity	
Method	**Corr.[a]**	**Method**	**Corr.[b]**	**Method**	**Corr.[c]**	**Method**	**Corr.[d]**
mCSM-Stability	0.69	mCSM-PPI	0.16	mCSM-NA	0.70	mCSM-lig	0.63
DUET	0.68	mCSM-PPI2	0.42				
DynaMut	0.70	BeAtMuSiC	0.28				
SDM2	0.61	MutaBind	0.41				
STRUM	0.79	FoldX	0.12				
PopMuSiC 2.1	0.63	MMPBSA	0.19				
CUPSAT	0.78						
Eris	0.75						
INPS-MD/3D	0.72						

[a]Pearson's correlation when evaluated on blind-test sets derived from the ProTherm database
[b]Kendall rank correlation coefficient on 1007 single-point mutations from CAPRI (T55)
[c]Pearson's correlation on 331 single-point mutations from 38 protein–nucleic acid complexes
[d]Pearson's correlation on 763 single-point mutations from 200 protein–ligand complexes

The mCSM platform is the only available approach to consider all possible molecular effects and has therefore formed the central component of our mutational analysis pipeline. All mCSM

Platform tools are available freely as websites compatible with most web-browsers, but Google Chrome is recommended. A summary of these methods and links to access them is described in Table 3.

Table 3
Computational tools available in the mCSM platform

mCSM tool	Type	Function
Arpeggio[a]	Protein interaction	Calculates 13 different types of interactions between atoms including hydrogen bonds, halogen bonds, carbonyl interactions, and others.
MTR-Viewer[b]	Missense tolerance	A measure of a gene's regional tolerance to missense variation.
mCSM-Stability[c]	Stability	Predict the effects of a mutation on the overall protein stability
SDM2[d]	Stability	Predicts the change in protein stability due to a single mutation using conformationally constrained environment-dependent amino acid substitution tables.
DUET[e]	Stability	Uses mCSM-Stability and SDM2 in order to create a consensus prediction the effects of a mutation on protein stability
DynaMut[f]	Flexibility	Looks to predict the effects of a mutation on protein stability, flexibility, and dynamics
mCSM-PPI[g]	Protein interaction	Predicts the effects of a mutation within a specified protein on its impact with overall protein–protein interactions.
mCSM-PPI2[h]	Protein interaction	Creates a similar prediction to PPI but incorporates the effects of mutations on interresidue noncovalent interaction network using graph kernels, evolutionary information, complex network metrics, and energetic terms.
mCSM-DNA[i]	Protein interaction	Predicts the impact of mutations on the protein interaction with DNA.
mCSM-NA[j]	Protein interaction	Predicts the impact of mutations on the protein interaction with nucleic acids, and uses pharmacophore and information about nucleic acid properties.
mCSM-Lig[k]	Protein interaction	Predicts the effects of single-point mutations on the stability of a protein–ligand complex.

[a] http://biosig.unimelb.edu.au/arpeggioweb/
[b] http://biosig.unimelb.edu.au/mtr-viewer/
[c] http://biosig.unimelb.edu.au/mcsm/stability
[d] http://marid.bioc.cam.ac.uk/sdm2
[e] http://biosig.unimelb.edu.au/duet/
[f] http://biosig.unimelb.edu.au/dynamut/
[g] http://biosig.unimelb.edu.au/mcsm/protein_protein
[h] http://biosig.unimelb.edu.au/mcsm_ppi2/
[i] http://biosig.unimelb.edu.au/mcsm/protein_dna
[j] http://biosig.unimelb.edu.au/mcsm_na/
[k] http://biosig.unimelb.edu.au/mcsm_lig/

3 Methods

3.1 Predicting and Analyzing Structural and Biophysical Effects of Mutations Using the mCSM Platform

The mCSM methods can be categorized by purpose. As shown in Fig. 1, methods are chosen depending on interactions made, and what structural information is available. Below we discuss how each type of predictor can be used and interpreted.

- The user should choose the appropriate tools based on what information is available on their protein of interest (Fig. 1).
- In general, each mCSM tool requires a wild-type protein file, in the PDB format, and the single-point mutation or a list of mutations. Some tools may require additional specific information; Table 4 shows the inputs required for each tool. **Notes 1** and **2** highlight some common issues with the submission inputs.

3.2 mCSM Platform Output

The results of Arpeggio are shown in Fig. 2.

3.2.1 Arpeggio

- After submitting a job, an overview of the type and number of atomic interactions within the protein is shown (Fig. 2a). Arpeggio calculates all types of molecular interactions (Table 5), which are displayed and downloadable along with a visual representation of the atomic contacts overlaid on the protein structure (Fig. 2b).
- The number of each interaction/contact and PyMOL session files can be downloaded for a more detailed analysis.

3.2.2 MTR-Viewer

Gene Viewer

- The MTR gene viewer [5] results page (Fig. 3) shows predicted MTR scores in an interactive line graph with a control panel which allows users to adjust the window size and the ethnicity for MTR estimates. A line graph (Fig. 3a) displays regions that have high variation, low-MTR scored; those in red are most likely to be pathogenic. Any ethnicity-specific MTR scores are shown in blue on the line graph.
- The first lollipop plot (Fig. 3b) shows observed missense (yellow) and synonymous (green) variations based on gnomeAD.
- If the gene of interest is a ClinVar pathogenic gene, their pathogenic (red) and benign (blue) missense variants are displayed under the gnomeAD lollipop plot (Fig. 3c).
- Users can browse results of alternative-transcript (Fig. 3d) of the given query if available.

Variant Query

- The variant query result page (Fig. 4) shows MTR scores for each user-supplied missense variant, providing the estimated regional intolerance. Low MTR scores indicate stronger purifying selection within the population. Users can also press "view" next to a variant to show its position within its gene transcript.

Table 4
Information required to run each mCSM program

		Inputs	
mCSM tool	**Task**	**Step 1**	**Step 2**
Arpeggio	Calculate	Molecule in PDB format or PDB accession code.	Select desired interaction calculation. You can select any (including multiple) part of the PDB file using the syntax: /1/2/3 Where: 1. Chain ID. 2. Residue number. 3. Atom name.
MTR-Viewer	Gene Viewer	Gene, ensembl ID, or Refseq ID	Select window size and overlay sub-population
	Variant Queries	Variants as GrCh37 genomic coordinates.	
mCSM-Stability, mCSM-PPI, mCSM-DNA	Prediction	Wild-type protein file in PDB format. For mCSM-PPI and mCSM-DNA, the structure of the complex in PDB format is required.	Single mutation (code and mutation chain), file with a list of mutations and its respective chains or code of residue and the mutation chain.
SDM2	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain) or residue/position code and the mutation chain.
DUET	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain)
DynaMut	Analysis	Wild-type protein structure in a PDB format or PDB accession code.	The selection of a Force Field and email (optional field).
	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field).
mCSM-PPI2	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field).
	Analysis	The structure of the complex in PDB format or corresponding PDB accession code.	Mutation details (alanine scanning or saturation mutagenesis) and email (optional field).
mCSM-NA	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and the selection of the Nucleic Acid Type.
mCSM-Lig	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) and ligand information (three-letter ligand ID and estimated wild-type affinity).

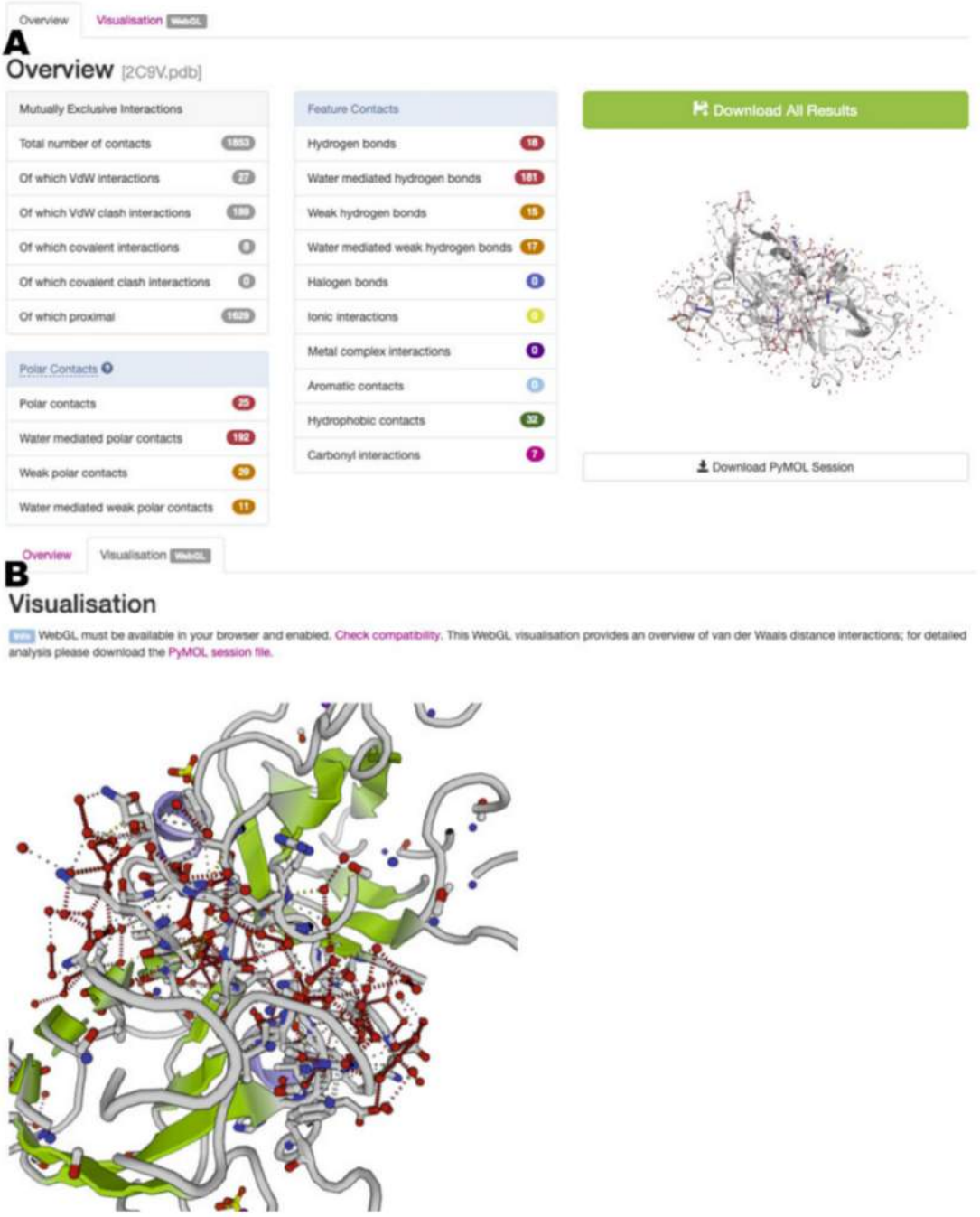


Fig. 2 Output of the Arpeggio tool. (**a**) Overview of the output for the inputted protein including the different types of interactions. (**b**) Visualization of the interactions shown on a protein structure

Table 5
Atomic interactions calculated by Arpeggio

Atomic interaction	Description	Arpeggio class	Bond energy (kJ/mol)
Van der Waals (dipole)	Permanent, induced and instantaneous dipoles	VWD	1–9
Hydrophobic	Between aliphatic and aromatic atoms	Hydrophobic	4–12
Hydrogen bond	Between carboxyl, amide, imidazole, guanidine, amino, hydroxyl and phenolic groups	Hydrogen bonds, weak hydrogen bond, polar contacts, halogen bonds, carbonyl interactions	8–40
Pi interactions	From/to rings	Aromatic contacts	6–70
Electrostatic	Between carboxyl and amino groups	Ionic interactions, metal complex	42–84

3.2.3 mCSM-Stability/PPI/DNA

The impact of mutations on protein stability, protein–protein binding affinity, and protein–DNA affinity can be predicted by mCSM-Stability, mCSM-PPI, mCSM-DNA with three types of prediction; single, multiple and systematic mutation.

Single Mutation

- If the single mutation option is selected in one of the tools within the mCSM platform, it will be shown on a results page after processing. This information includes the predicted value changes (protein stability, protein–protein interaction, protein–DNA interaction) as measured by the change in Gibbs Free Energy $\Delta\Delta G$ kcal/mol (Fig. 5), which is classified as highly destabilizing ($\Delta\Delta G \leq -2$ kcal/mol), destabilizing (-2 kcal/mol $< \Delta\Delta G < 0$ kcal/mol), stabilizing (0 kcal/mol $\leq \Delta\Delta G < 2$ kcal/mol), or highly stabilizing ($\Delta\Delta G \geq 2$ kcal/mol).
- If the structure of a complex is submitted to mCSM-Stability, it will calculate the predicted change in stability of the entire complex. It is therefore often advisable to also run predictions on a PDB file containing the protomer chain alone.
- For mCSM-PPI and mCSM-DNA, for mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there are fewer mutations located further away than 12 Å in the datasets used to train the methods.
- Also shown is an interactive 3D visual representation of the uploaded PDB file (Fig. 5a, right).

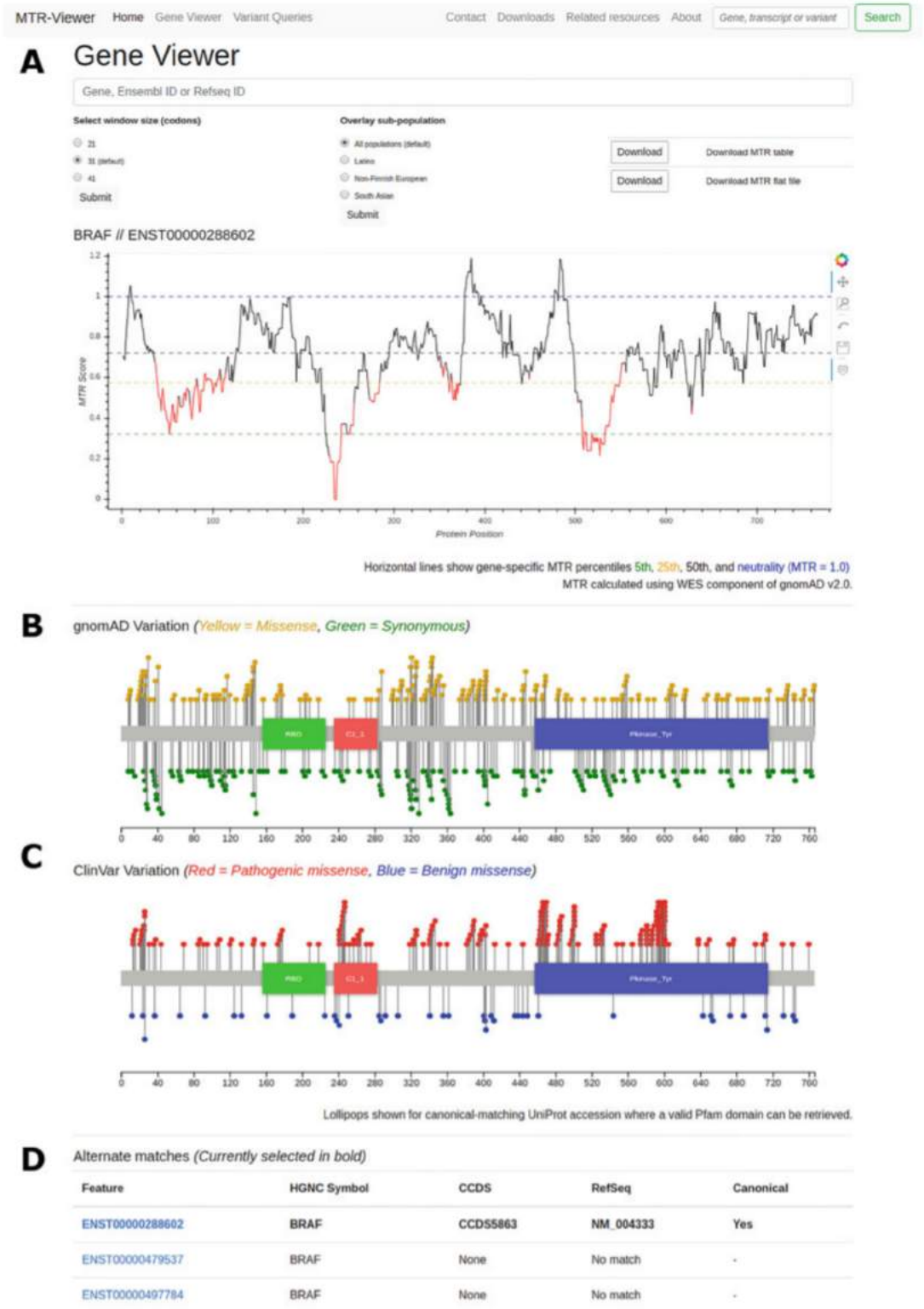
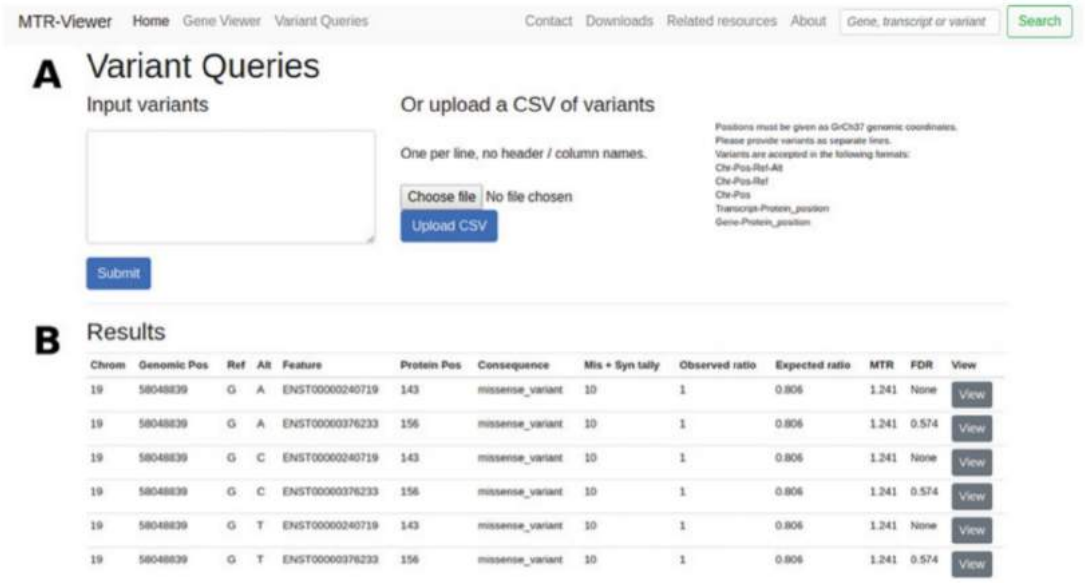


Fig. 3 The MTR Gene Viewer result page. (**a**) The line graph shows MTR scores in red for variations distant from neutrality across the transcript according to selected window size (codons) and subpopulation option. (**b**) The lollipop plot shows observed gnomeAD variation in yellow and green for missense and synonymous variation. (**c**) The second lollipop plot displays pathogenic (red) and benign (blue) missense variants based on ClinVar annotation. (**d**) The alternate transcripts can be shown in a table with RefSeq ID



Chrom	Genomic Pos	Ref	Alt	Feature	Protein Pos	Consequence	Mis + Syn tally	Observed ratio	Expected ratio	MTR	FDR	View
19	58048839	G	A	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	A	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View
19	58048839	G	C	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	C	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View
19	58048839	G	T	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	T	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View

Fig. 4 MTR Variant Queries result page. Calculated results and information for the given input variants (or a CSV). User can check the details through MTR Gene Viewer by clicking on the view button

Multiple or Systematic

- If the option for inputting a list of mutations or systematic was used to analyze the PDB file, then after processing, results will be shown in tabulated form (Fig. 5b), including mutation specific information such as the residue solvent accessibility (RSA), as well as the predicted ΔΔG.
- Each result is also classified, using the predicted ΔΔG value, as highly destabilizing, destabilizing, stabilizing, or highly stabilizing.
- Users can search the result table or download results into a tab-separated text file.

3.2.4 SDM

SDM uses environment-specific amino acid substitution tables [38] and structural features including residue depth [15] and packing density to predict the impact of mutations on protein stability. The result page of single and list mutation is as follows.

Single Mutation

- The single mutation result page (Fig. 6a) provides predicted protein stability changes (ΔΔG), in addition to structural information implemented in SDM including secondary structure, RSA, residue depth and residue occluded packing density (OSP), sidechain–sidechain hydrogen bond (HBOND_SS), sidechain–main chain amide hydrogen bond (HBOND_SN), and sidechain–main chain carbonyl hydrogen bond (HBOND_SO). The integrated 3D viewer also shows the

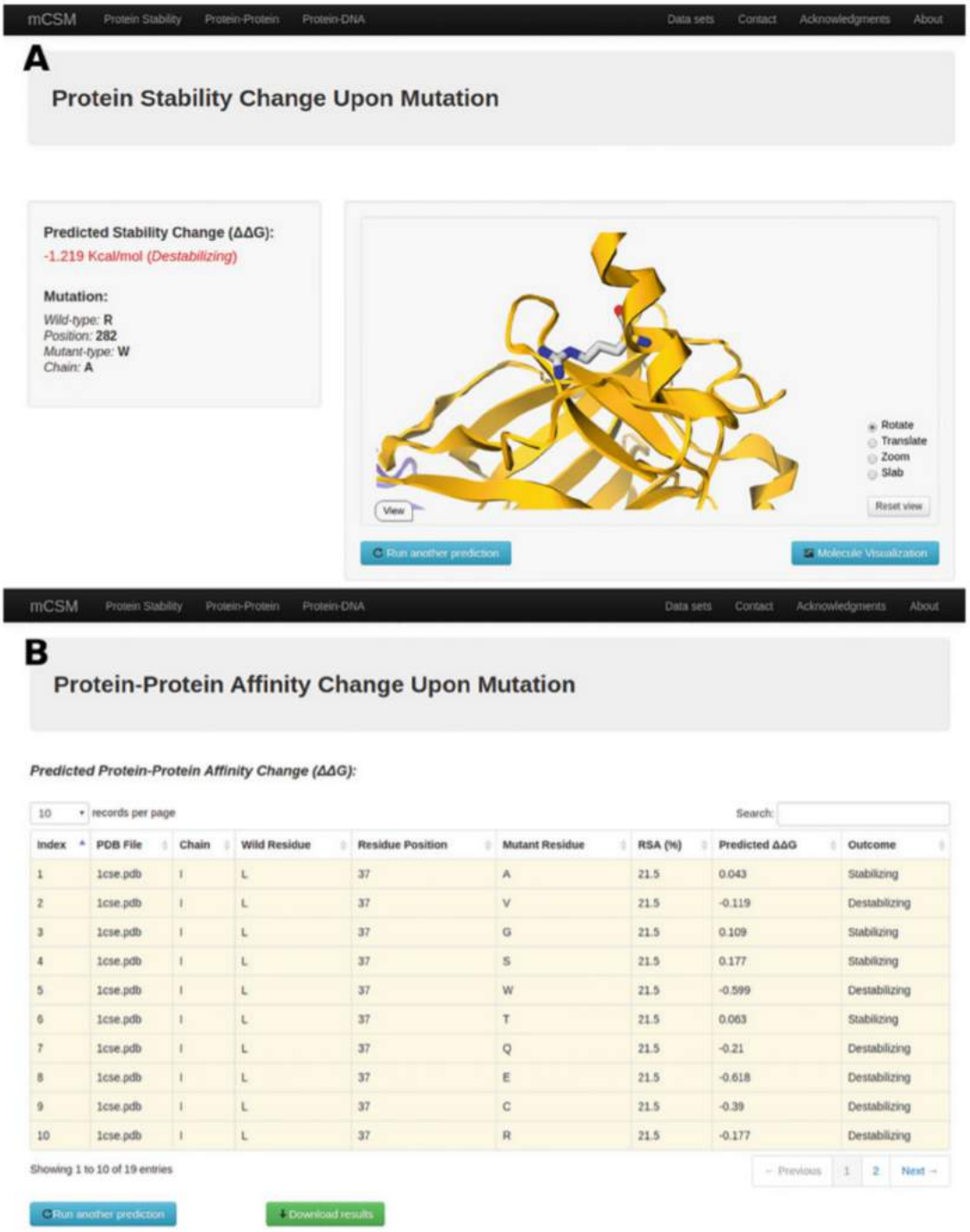


Fig. 5 Result pages for mCSM-Stability, mCSM-PPI and mCSM-DNA. (**a**) mCSM-Stability (single mutation) and (**b**) mCSM-PPI (multiple/systematic mutation). (**a**) The single prediction for example mCSM-Stability page supports 3D interactive viewer for structural analysis. (**b**) The results and information from multiple/systematic prediction for example mCSM-PPI are shown in a table

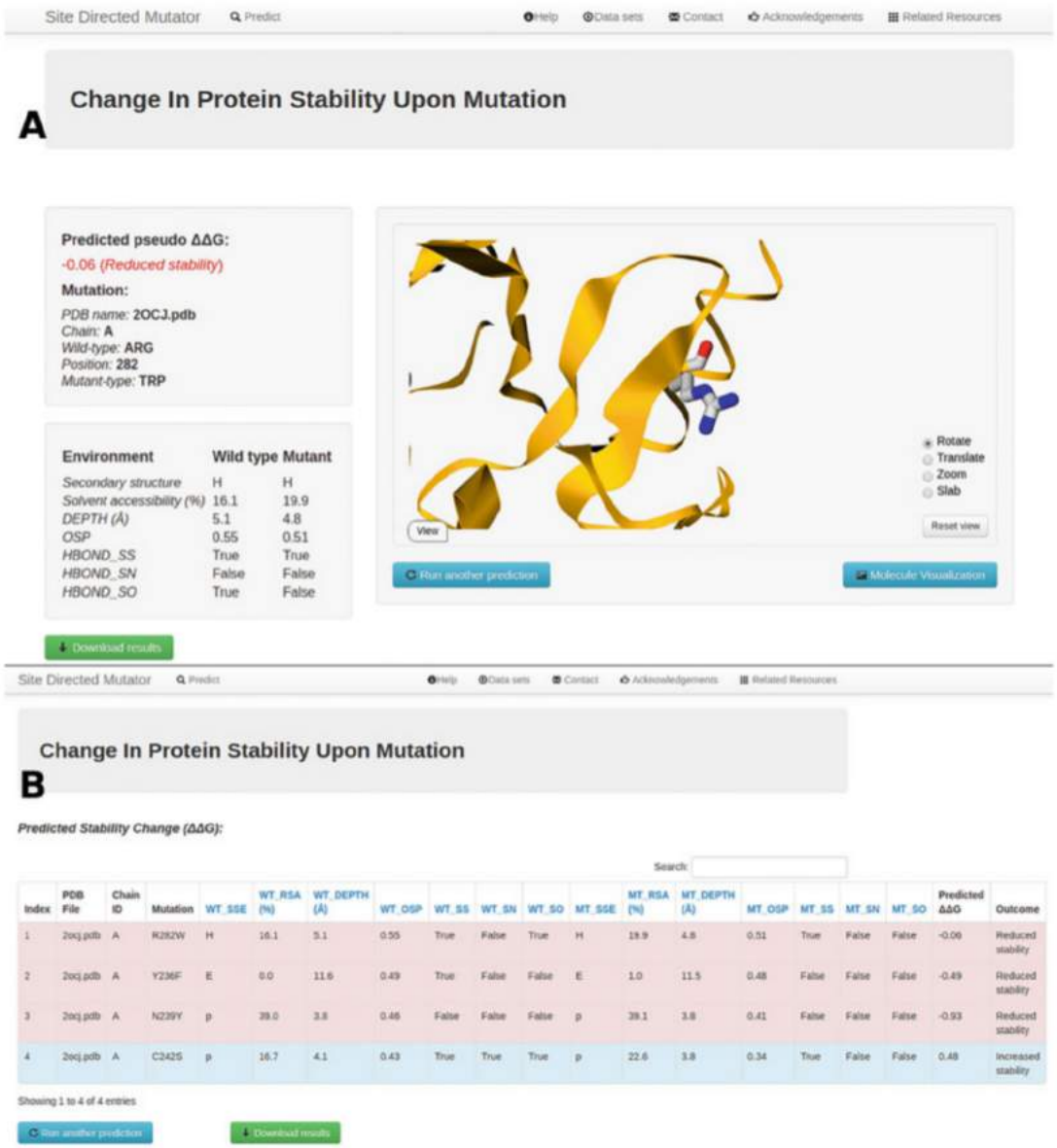


Fig. 6 SDM prediction results for single and list prediction. (**a**) The single prediction displays the predicted ΔΔG with information used on the left panel and 3D structure in a ribbon (protein) and a stick (wild-type amino acid) representation. (**b**) The list prediction gives detailed structural information and predicted ΔΔG in a tabulated form highlighted according to stabilizing (blue) and destabilizing (red) mutation

structure and its wild-type amino acids in ribbon and stick representation.

- Stability changes (ΔΔG) are shown in red with a negative sign if the mutation is predicted to be destabilizing, and in blue with a positive sign if the mutation is predicted to be stabilizing.

Multiple Mutations

- The predicted SDM ΔΔG for a given mutation list is displayed in a tabulated format (Fig. 6b) with their structural features. Users can download all mutant PDB structures and their predicted values in individual files.

3.2.5 DUET

Single Mutation

- The DUET result page (Fig. 7a) provides the predicted stability changes (ΔΔG) with integrated features such as secondary structure and stability changes from mCSM and SDM. While DUET refers to both mCSM and SDM scores, the prediction result can vary between the two methods.
- In the structure viewer (Fig. 7a right), the wild-type amino acid is shown in stick form and users can download the corresponding mutant structure file in PDB format.

Systematic Mutations

- With the systematic prediction (Fig. 7b), users can examine the predicted changes in protein stability using DUET, mCSM, and SDM for all nineteen possible mutations at a given residue position.
- The predictions and the structural information used to calculate the DUET scores are displayed in a downloadable table.

3.2.6 DynaMut

Users can use DynaMut to assess the impact of mutations on protein dynamics and stability with single and list mutation prediction.

Single Mutation

- The results of mutational effects on protein dynamics and stability are shown in Fig. 8a: ΔΔG predictions, interatomic interactions, deformation and fluctuation analysis.
- The ΔΔG prediction page provides predicted values from normal mode analysis (NMA)-based prediction (ΔΔG ENCoM), vibrational entropy energy changes ($\Delta\Delta S_{Vib}$ ENCoM), and other structure-based stability predictions (ΔΔG mCSM, ΔΔG SDM, ΔΔG DUET). Users can visually assess mutational effects on protein flexibility which is colored on the protein structure by vibrational entropy (Fig. 8b) for the region gaining (red) or losing (blue) flexibility. This 3D representation can be downloaded into a Pymol session, high resolution image and CSV file.
- Through the interatomic interactions tab, users can compare molecular interactions between wild-type and mutant structures. The PDB structure with interatomic interactions can be retrieved as a Pymol session file.
- The mutational effects on protein dynamics are shown in the deformation and fluctuation tab. Users can evaluate changes in the amount of local flexibility and atomic fluctuation upon mutation in 3D visual representation; results are downloadable as a CSV file and a Pymol session file.

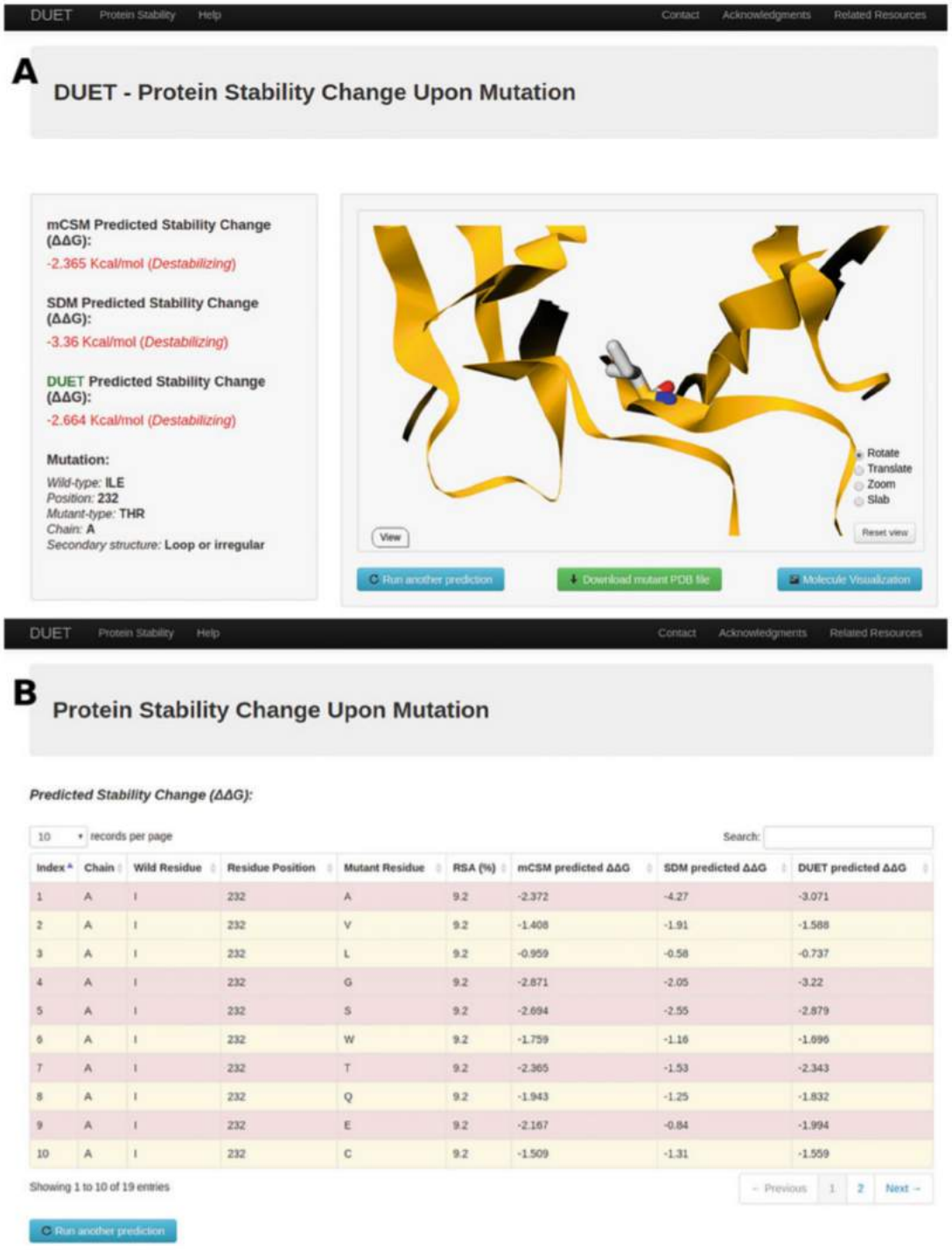


Fig. 7 DUET result pages for single and systematic prediction. (**a**) The single prediction result of DUET shows predicted ΔΔG across SDM and mCSM-Stability with mutation details. (**b**) Systematic prediction results including ΔΔG from DUET, SDM and mCSM-Stability and relative solvent accessible area of wild-type structure

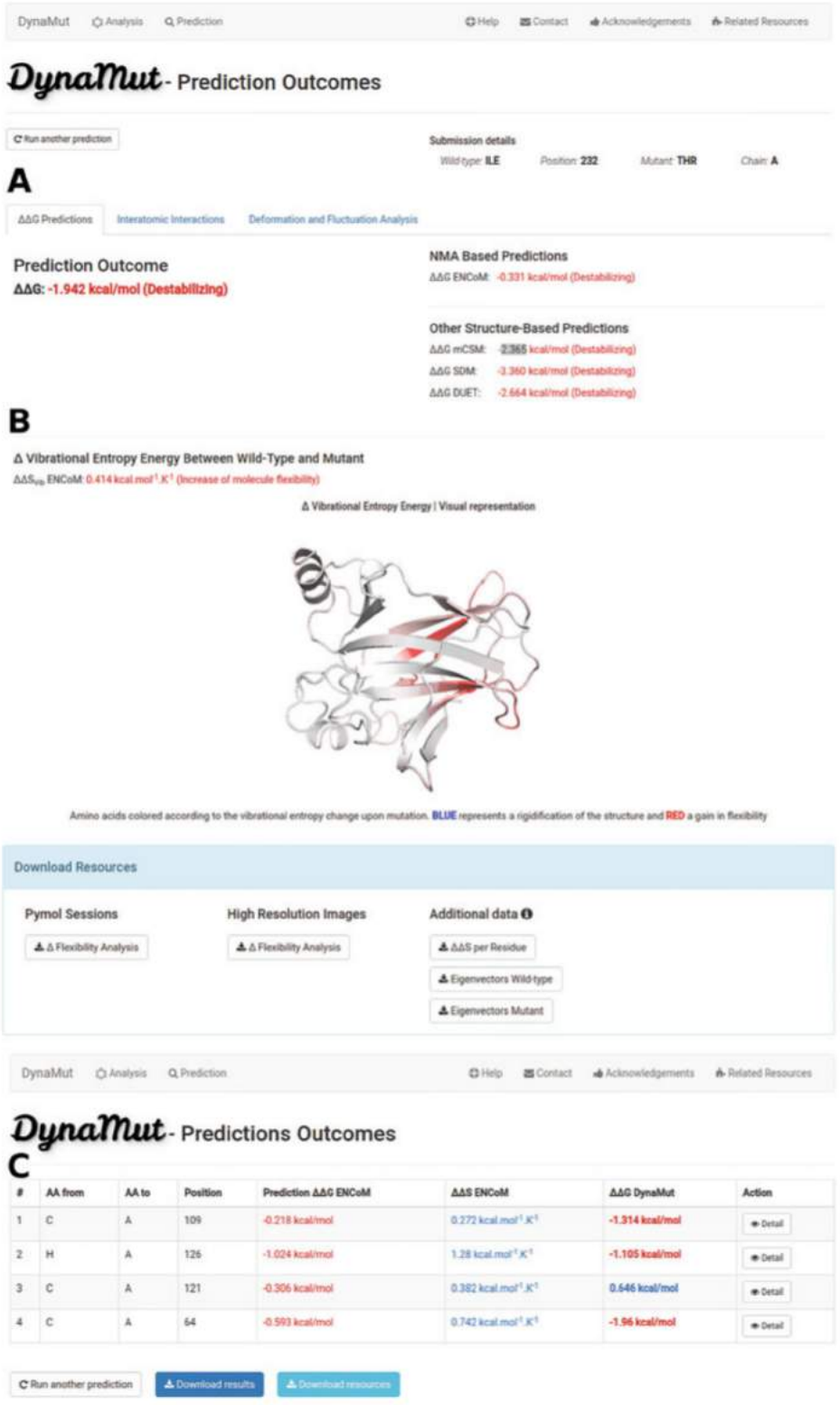


Fig. 8 DynaMut result pages. The single prediction shows predicted DynaMut ΔΔG (**a**, left) and predicted protein stability (ΔΔG) from mCSM-Stability, SDM and DUET and flexibility changes (ΔΔG ENCoM). Users can check vibrational energy changes upon mutation in the panel B. For a multiple mutation, (**b**) list prediction result page shows predicted DynaMut ΔΔG and links to access the corresponding single prediction in table

Multiple Mutations

- For a given mutation list, DynaMut gives all predicted values, including $\Delta\Delta G_{Stability}$ENCoM, $\Delta\Delta S_{Vib}$ ENCoM, and $\Delta\Delta G_{Stability}$ DynaMut, in table format (Fig. 8c). A more detailed analysis is available through the single prediction page of each mutation by clicking on the "Detail" button.

3.2.7 *mCSM-PPI2*

mCSM-PPI2 supports two types of protein–protein affinity prediction: mutation prediction and binding analysis. Mutation prediction gives predicted protein–protein affinity changes based on a given protein–protein complex and the mutation information. Binding analysis considers interface residues within 5 Å from different chains in the complex structure for alanine scanning and saturation mutagenesis.

Single Mutation

- mCSM-PPI2 displays predicted binding affinity changes (ΔΔG) upon mutation in two classes, destabilizing and stabilizing. Mutation details such as the distance to the interface from the given mutation position are also shown (Fig. 9).
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- Users can assess the mutational impact in atomic/residue level through a 3D interactive viewer and a 2D graph. The molecular viewer provides Arpeggio inter/intra interactions for wild-type and mutant structures and the interaction changes between wild-type and mutant allows for investigation of the relationship between nonbonded interaction and protein–protein affinity. For residue-level analysis, the 2D graph can be used to study interresidue interactions of wild-type and mutant in a simple and user-friendly representation.

List Mutation

- For multiple mutation analysis, the result page tabulates predicted ΔΔG with mutation details. Users can access detailed results of each mutation through the single mutation result page and download all entries as a CSV file.

Alanine Scanning

- To identify residues with a greater contribution to the energy of binding (hot-spot) at the interface of interaction, alanine scanning can be used by predicting protein–protein binding affinity changes upon mutations to alanine across all identified interface residues. The predicted ΔΔG values are displayed in table, bar chart, and 3D viewer (Fig. 10a).
- Users can assess the effects of alanine mutation on the interface residues through a bar graph and 3D viewer colored in red and blue for destabilizing and stabilizing mutations, respectively.

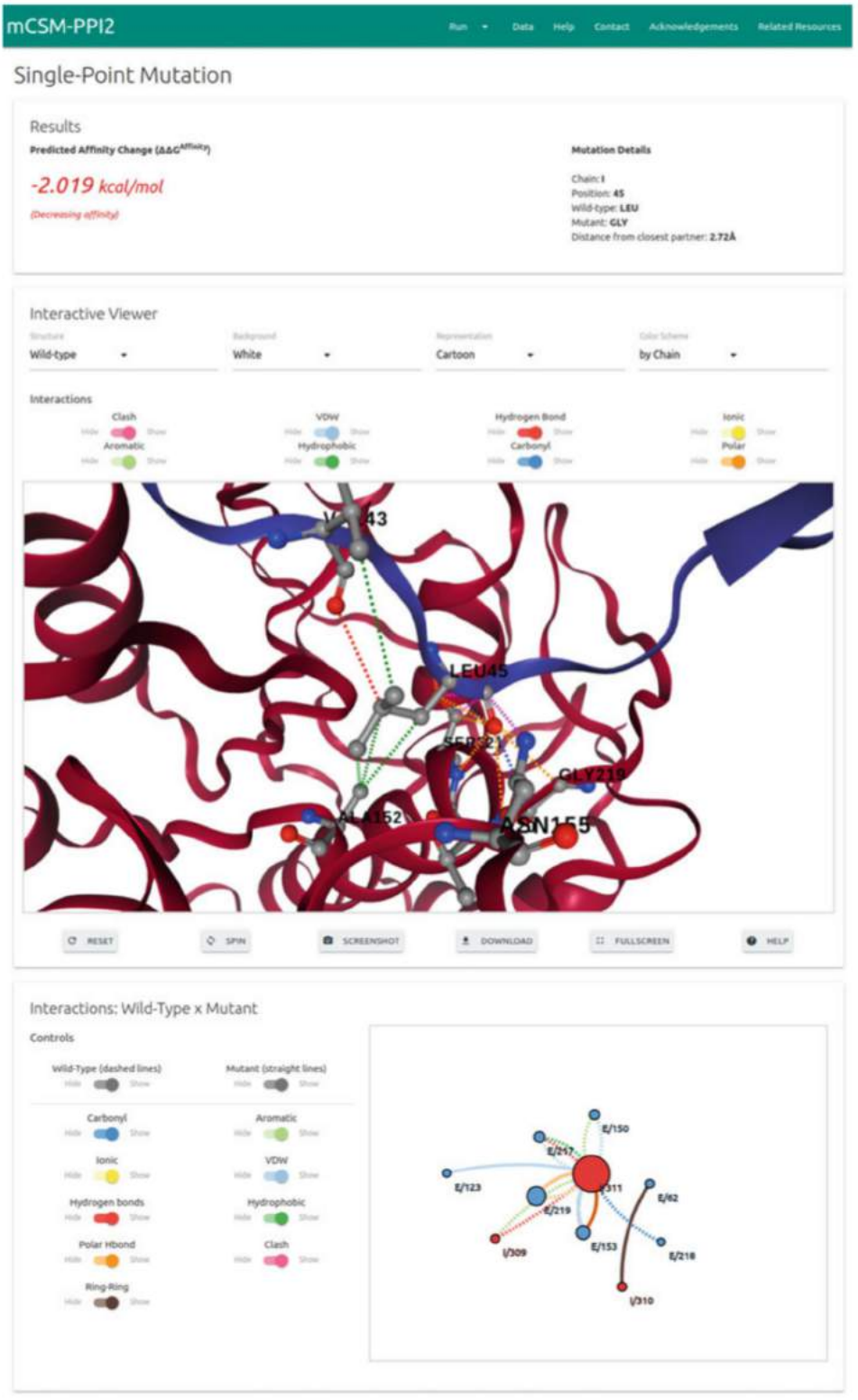


Fig. 9 mCSM-PPI2 single prediction result page. The predicted ΔΔG is shown along with two interaction viewers: 3D interactive molecule viewer for atomic interaction analysis and 2D diagram for residue-level interaction analysis

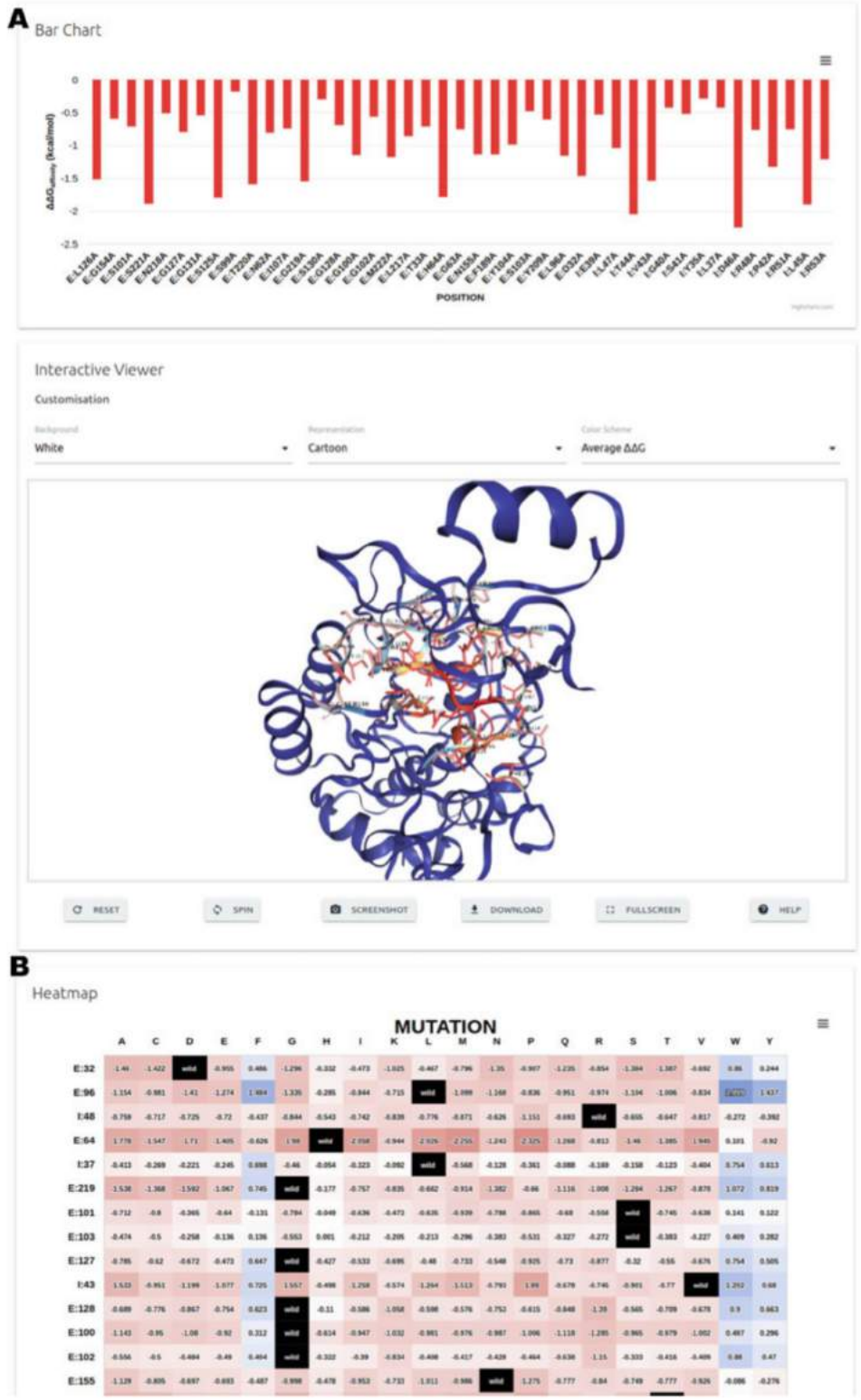


Fig. 10 mCSM-PPI2 interface scanning result pages. The result pages of (**a**) alanine scanning and (**b**) saturation mutagenesis provide a bar chart and a heatmap colored by predicted ΔΔG and average predicted ΔΔG from the nineteen possible mutations, respectively

Saturation Mutagenesis

- The saturation mutagenesis provides the most exhaustive prediction, showing predicted ΔΔG for all identified interface residues when they are changed into nineteen different amino acids. The results are shown in table, heatmap, and 3D molecule viewer, and the interface residues of the 3D viewer are colored by the average ΔΔG of all mutations for each residue.

3.2.8 *mCSM-NA*

Single Prediction

- The predicted protein–nucleic acid affinity changes on a given structure are shown (Fig. 11a) with other properties such as the type of nucleic acid, solvent accessibility of wild-type protein, and predicted mutational effects from mCSM-Stability.
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- The molecule visualization panel shows the protein–nucleic acid complex with the wild-type amino acid, and the mutation as a stick representation. mCSM-NA allows users to further investigate inter/intraresidue interactions by downloading Pymol session file.

List Mutation

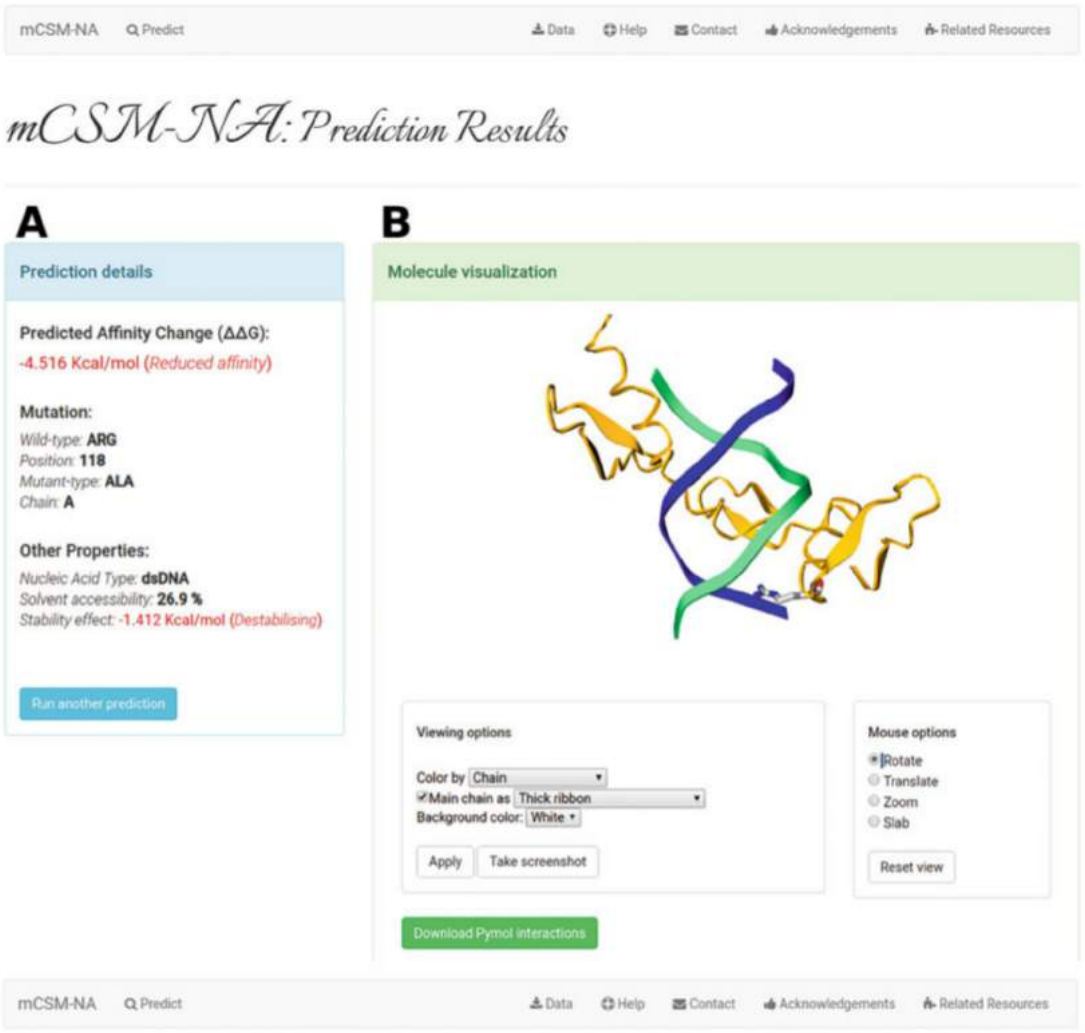
- mCSN-NA provides predicted protein–nucleic acid affinity changes, wild-type RSA, and mutation information for a given list of mutations in a table which is also downloadable in TSV format.

3.2.9 *mCSM-lig*

- mCSM-lig predicts affinity changes (log affinity fold) between a protein and its ligand upon mutation (Fig. 12a) using additional information such as the closest distance between wild-type residue and ligand and the protein stability change (Kcal/mol) from DUET. The stabilizing and destabilizing mutations are shown in positive and negative values respectively.
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- The wild-type amino acid and ligand are shown in stick and sphere representations in 3D molecule viewer, respectively.

3.3 Identification of Driving Molecular Consequences

The outputs of the predictive tools described above provide the basis for an initial heuristic examination. When trying to interpret the molecular consequences of a specific variant, it is important to remember that phenotypic outcomes are often the result of the



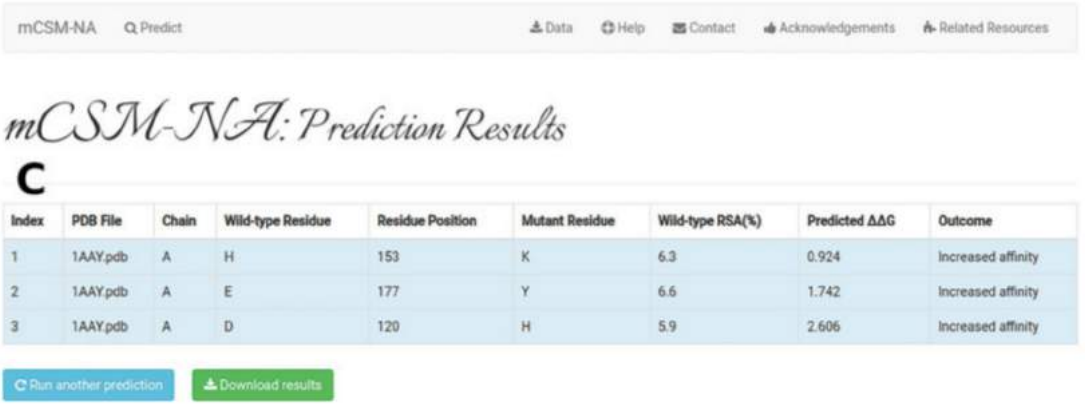


Fig. 11 mCSM-NA result pages for single and list mutation prediction. In the single prediction result page, predicted protein–DNA affinity changes and mutation information are displayed in the prediction details (**a**) and the 3D viewer shows protein–DNA complex and wild-type amino acid in a ribbon and stick representation (**b**). The results of list prediction are shown in a tabulated form (**c**) and users can save the results in a TSV format

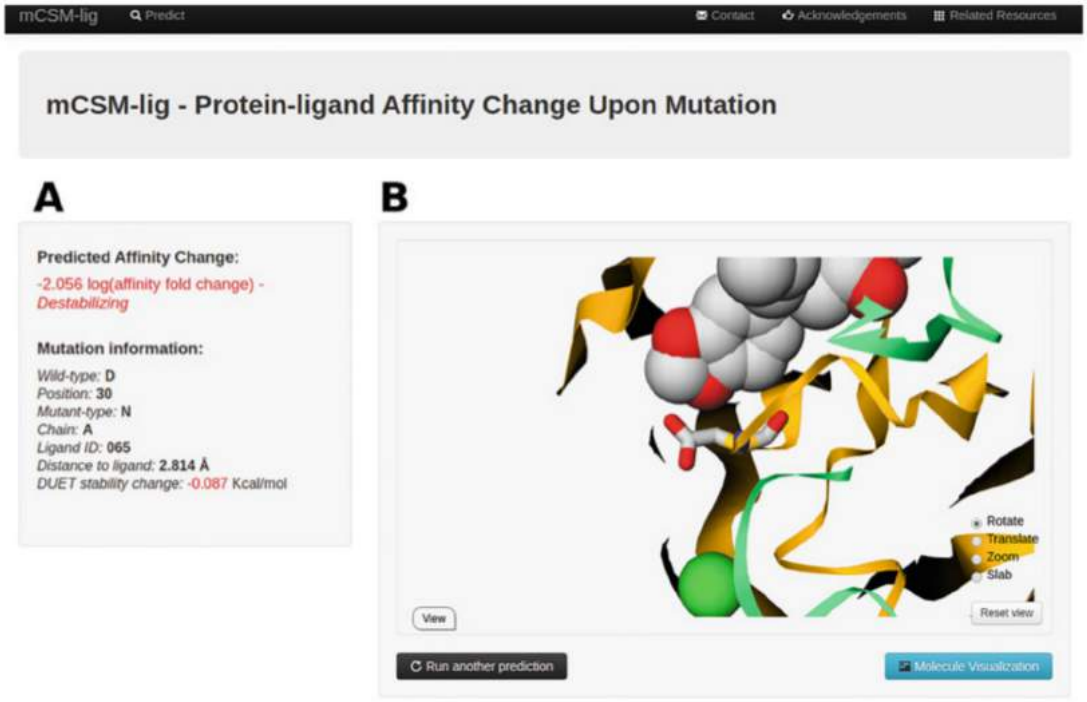


Fig. 12 mCSM-lig result page. (**a**) The predicted affinity change between protein and ligand upon mutation is shown in logarithm scale. (**b**) The protein and ligand are displayed in 3D viewer with a ribbon (for protein), a stick (for wild-type amino acid), and a sphere (for ligand) representation

combination of multiple molecular changes. For coding mutations, we initially ask ourselves three questions:

1. Is the mutation within 5 Å of an interface? If so, is the mutation more likely to disrupt the interaction ($\Delta\Delta G < \pm 0.5$ kcal/mol) based on the corresponding mCSM output (e.g., mCSM-PPI, mCSM-DNA, mCSM-NA, mCSM-Lig)? If the mutation is further than 12 Å away, it is less likely to disrupt the interaction directly, so the mCSM predictions are less reliable.
2. Is the mutation likely to disrupt protein folding and stability? mCSM-Stability, SDM, DUET, and DynaMut provide insight into this, with mutations leading to $\Delta\Delta G < \pm 0.5$ kcal/mol more likely to have a significant biological effect. Mutations at buried residues are more likely to have a larger effect on protein stability.
3. Is the mutation a special case that is more likely to lead to disruption of the protein due to unique geometry restraints of the residues (*see* **Notes 3** and **4**)?

To more exhaustively explore how mutations in a protein lead to a phenotype, and to identify those molecular features that best

capture the driving of the molecular mechanisms, an investigation into the performance of each inputted feature should be conducted in order to construct the highest performing predictive model.

A more robust method for selecting which features are most informative can be performed using feature selection in R, a statistical programming language. While R is powerful enough itself to create classification models, we can also use it to measure which features from our predictive tools' output are most effective in stratifying mutations. Two effective approaches are:

1. A random forest classification algorithm to measure feature importance using a set of mutations with known class labels (e.g., pathogenic/nonpathogenic, deleterious/nondeleterious).
2. The Boruta Algorithm performs permutations of the data to statistically compare each feature's importance with that attainable at random, and uses this to eliminate uninformative features. The package in R provides a graphical output using boxplots.

Features that score highly provide evidence that the molecular consequence that they measure is relevant to how mutations lead to the phenotype of interest. The algorithm can also highlight correlation between features. When two or more features are highly correlated and are likely measuring the same information, only one should be used in subsequent predictive model development to remove redundancy, minimize noise and avoid bias from weighting a model in favor of a particular attribute. The model should also have the fewest possible features that perform best. Using too many features may generate a model that performs accurately on training data but cannot be generalized to real-world data.

3.4 Machine Learning Phenotypes: Building a Predictive Classifier

An initial understanding of molecular mechanisms imparted by disease-causing mutations is a crucial step toward establishing a genotype–phenotype correlation. However, manual analysis of different results can often miss underlying, statistically significant relationships among different mutational measurements, which can help relate them to the phenotype. Machine learning, and in particular supervised learning, addresses this issue by providing a set of tools for the efficient analysis of labeled data (e.g., experimentally characterized mutations) in order to derive a model that describes a phenomenon, aiming for generalization (applying it to unseen data). The identification of patterns and associations within the data will further help the predictive model establish a distinction between mutations within the same gene leading to different phenotypes, and hence the development of an effective predictive tool that can be used to interpret novel clinical variants.

Here, our goal is to build a machine learning classifier to distinguish between pathogenic vs. nonpathogenic mutations in a

given gene. Multiple steps are required to obtain a nonbiased, accurate predictor:

1. Dataset curation: Machine learning algorithms require a well-curated dataset. In a supervised machine learning approach, all data labels (here, pathogenic or nonpathogenic for each mutation) must be known in order to enable correlations to be assessed between labels (e.g., phenotypes) and features/properties used as evidence to represent each data point (e.g., mutations). The quality of a classifier directly depends on the quality of the data used to build it, so accurate clinical sources are required to justify labeling mutations as pathogenic or nonpathogenic. In this case, generally, nonpathogenic variants can be curated from population variant databases such as GnomAD, usually taking into account frequent mutations. Even common variants, however, may still be linked to a disease, especially if it is a weakly penetrative mutation or recessive condition, which would add noise to the data set and thus complicate the task of building a general predictive model. In situations where other biologically relevant information is present, such as cellular fitness cost, it is essential that this type of information is present for every mutation in a dataset, as a supervised algorithm cannot handle missing data labels. The initial dataset should contain a representative set of mutations within all phenotype classes (pathogenic and nonpathogenic), and ideally, present a balanced number of instances between classes, to minimize biases toward overrepresented classes in the resultant model. More details on metrics used to evaluate the performance of predictive models on an imbalanced dataset are discussed below.
2. Feature generation: The feature generation stage is crucial as it provides descriptive information about each mutation, to be used by the learning algorithm to finally classify the phenotype of a mutation. As described above, features can encompass a diverse range of mutational information:
   (a) Protein stability and dynamics (mCSM-Stability, DUET, SDM, Dynamut).
   (b) Protein functional changes such as changes in affinity for other proteins (mCSM-PPI2), nucleic acids (mCSM-NA), and ligands (mCSM-lig).
   (c) At the residue level, changes in protein pharmacophore and local residue environment such as changes in interatomic interactions (Arpeggio) are also important, as some mutations at the same locus can have different phenotypes.
   (d) Sequence-level predictors (SIFT, Polyphen, SNAP2).

(e) Evolutionary-based predictors (ConSurf), population based mutational tolerance (MTR-Viewer), as well as amino acid substitution matrices (e.g., PAM30, BLOSUM62, PSSM) offer added information on the likelihood of one mutation to change into another.

Feature generation is directly dependent on the wild-type biological functions of the protein, which is why an understanding of the biological relevance is important at the very beginning of this process.

3. Training and Testing sets: The data collected must be divided into training and testing sets to assess the generalization power of a classifier, that is, its ability to correctly predict on new data, and to ensure that it has not been over- or undertrained. Data used to train the model should be different, nonredundant, from the data used to test the model. It is common practice to divide the original dataset into Training and Test sets at the start of learning. For small datasets, a large proportion of the data may need to be segregated into the Test set to provide sufficient data to accurately measure performance of the trained model. This can be done in a bootstrapping procedure or through cross-validation, when the original data set is divided into *k*-folds and each is taken iteratively as the test set while remaining data are used in training (*k*-fold cross-validation).
4. Feature selection: The features selected for training can strongly influence accuracy, so it is important to select only informative features, and eliminate irrelevant or nondiscriminative ones, which are a common source of noise. Feature selection can also help reduce overfitting and reduce training time, as it aims to generate simpler, more concise models. Feature selection methods provided in the Python machine learning library, Scikit-Learn [58], include univariate selection, feature importance, correlation matrix, and recursive feature elimination or addition. Alternatively, forward stepwise selection can be performed as a greedy heuristic in which features are included iteratively, one at a time, based on their individual performance contributions.
5. Machine learning platforms: Different tools have been developed for implementing machine learning. Some offer a graphical user interface (GUI), such as Weka [59], while some run as python packages through the command line, such as Scikit-Learn. Different packages for different programming languages offer similar algorithms and options to adjust the algorithm parameters according to specific tasks. The major classification algorithms we test are Naive Bayes, Decision Trees, K-Nearest Neighbor, Support Vector Machines, and Ensemble Classifiers. It is good practice to compare

representative algorithms of each class, provided that the algorithm is compatible with the dataset type. Within weka, this can be done automatically using the auto-weka function. In cases where the training set is unbalanced, oversampling or undersampling of the training data can be used to achieve a better representation of classes within the classification model-building stage, preventing model bias in always detecting the predominant class and achieving a false high performance.

6. Model validation: The primary tool in the validation of a model is the use of a nonredundant independent test set, also called blind test.

   Validation can be furthered using internal data testing such as k-fold cross validation, in which the dataset is divided into k subsets. One subset is used as a test set, while the remaining $(k - 1)$ subsets are used to train a model. The process is repeated k times, until all the data have been used in both training and test sets. The final model performance is calculated as the average of the performances of all k iterations. We will often vary k based on the size of the dataset. When the training set is small (e.g., ~200 data points), we may use leave-one-out validation, where k is equal to the size of the dataset. An important aspect when selecting predictive models is consistency in performance between the training and test sets. This usually indicates a robust model, within which discrepancies (e.g., a much higher performance on training than with the test set) might indicate overfitting.

7. Model evaluation: Several different evaluation metrics may be used for classification tasks. These are generally calculated on values obtained from a confusion matrix, which is a summary of the data points, and their actual and predicted phenotypes (Table 6).

8. From the distributions of data points within the matrix, descriptive metrics can be calculated:

   (a) accuracy (number of correct predictions: [(TP + TN)/TOTAL]),

   (b) precision (rate of correctly predicted positive instances from all assigned as positives: [TP/(TP + FP)],

Table 6
Description of a confusion matrix

Predicted value	Actual value Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

(c) recall (rate of correctly predicted positive instances from all real positive instances: [TP/(TP + FN)],

(d) *f*-score (a weighted average of recall and precision), and,

(e) Matthews correlation coefficient (MCC) a balanced measure between true positives and true negatives

$$\left[(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN}) / \sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}\right]$$

where TP = True positive; TN = True negative; FP = False positive; and FN = False negative.

Classifier performance can also be described graphically using a Receiver Operating Characteristic curve, which compares the TP Rate and TN Rate. The closer the area under the curve is to 1, the better the classifier performance.

These metrics should be used in a combinatorial fashion across all elements of training, test, and cross-validation stages to compare model performance during different stages of classifier optimization. When the dataset is imbalanced, balanced measures such as MCC should be prioritized, as other measures might bias for an overtrained model on the dominant dataset.

4 Notes

1. Often following curation, the distribution of number of pathogenic and benign mutations is unbalanced, which can affect efforts to build predictive tools using machine learning. Two approaches that can help include oversampling of the underrepresented class, or undersampling of the overrepresented class. Evaluation metrics that are less biased toward unbalanced classes, such as the Matthew's correlation coefficient, precision-recall curves, and Kendall correlations, should also be preferentially used.
2. The chain ID for the provided PDB file is a mandatory field for all the structure-based methods; blank characters are not allowed. It is possible that homology modeling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. Several tools exist to perform this task (e.g., http://www.canoz.com/sdh/renamepdbchain.pl).
3. Special cases: Mutations to and from prolines. Prolines are the only amino acid whose amino group is connected to the side-chain, which in the context of the peptide bond greatly limits torsional angles. The nature of this residue therefore needs to be taken into account while analyzing mutation effects. For instance, (1) mutations to prolines in the middle of alpha-

helices can introduce kinks, affecting local structure and (2) since prolines are commonly found in turns and loops, their substitution might interfere with the formation of secondary structures such as hairpins.

4. Special cases: mutations of positive-phi glycines. Similarly to prolines, positive phi glycines, while rare in experimental structures, deserve special consideration due to their torsional angles. Glycines are the only residues capable of adopting positive-phi angles. These glycines are usually conserved across evolution, meaning that mutations on positive-phi glycines, especially on loops and hairpins, tend to be destabilizing.

Acknowledgments

This work was supported by Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; and the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.].

References

1. Jatana N, Ascher DB, Pires DEV et al (2019) Human LC3 and GABARAP subfamily members achieve functional specificity via specific structural modulations. Autophagy:1–17. https://doi.org/10.1080/15548627.2019.1606636
2. Abayakoon P, Jin Y, Lingford JP et al (2018) Structural and biochemical insights into the function and evolution of sulfoquinovosidases. ACS Cent Sci 4(9):1266–1273. https://doi.org/10.1021/acscentsci.8b00453
3. Ascher DB, Cromer BA, Morton CJ et al (2011) Regulation of insulin-regulated membrane aminopeptidase activity by its C-terminal domain. Biochemistry 50(13):2611–2622. https://doi.org/10.1021/bi101893w
4. Portelli S, Phelan JE, Ascher DB et al (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8(1):15356. https://doi.org/10.1038/s41598-018-33370-6
5. Silk M, Petrovski S, Ascher DB (2019) MTR-Viewer: identifying regions within genes under purifying selection. Nucleic Acids Res 47(W1):W121–W126. https://doi.org/10.1093/nar/gkz457
6. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
7. Lucy G, Douglas EVP, Álvaro O-N et al (2014) An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. Human Molecular Genetics, 23(22):5976–5988. https://doi.org/10.1093/hmg/ddu321
8. Blaszczyk M, Harmer NJ, Chirgadze DY et al (2015) Achieving high signal-to-noise in cell regulatory systems: spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. Prog Biophys Mol Biol 118(3):103–111. https://doi.org/10.1016/j.pbiomolbio.2015.04.007
9. Jafri M, Wake NC, Ascher DB et al (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096

10. Pacitto A, Ascher DB, Wong LH et al (2015) Lst4, the yeast Fnip1/2 orthologue, is a DENN-family protein. Open Biol 5 (12):150174. https://doi.org/10.1098/rsob.150174
11. Pires DE, Chen J, Blundell TL et al (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
12. Albanaz ATS, Rodrigues CHM, Pires DEV et al (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
13. Casey RT, Ascher DB, Rattenberry E et al (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
14. Jubb HC, Pandurangan AP, Turner MA et al (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
15. Pandurangan AP, Ascher DB, Thomas SE et al (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
16. Sibanda BL, Chirgadze DY, Ascher DB et al (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355 (6324):520–524. https://doi.org/10.1126/science.aak9654
17. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46(W1): W127–W132. https://doi.org/10.1093/nar/gky375
18. Hnizda A, Fabry M, Moriyama T et al (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32 (6):1393–1403. https://doi.org/10.1038/s41375-018-0073-5
19. Andrews KA, Ascher DB, Pires DEV et al (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55 (6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
20. Usher JL, Ascher DB, Pires DE et al (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
21. Nemethova M, Radvanszky J, Kadasi L et al (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
22. Ramdzan YM, Trubetskov MM, Ormsby AR et al (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
23. Traynelis J, Silk M, Wang Q et al (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27 (10):1715–1729. https://doi.org/10.1101/gr.226589.117
24. Trezza A, Bernini A, Langella A et al (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58(12):5320–5328. https://doi.org/10.1167/iovs.17-22158
25. Ascher DB, Spiga O, Sekelska M et al (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. Eur J Hum Genet 27(6):888–902. https://doi.org/10.1038/s41431-019-0354-0
26. Soardi FC, Machado-Silva A, Linhares ND et al (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2:7. https://doi.org/10.1038/s41525-017-0009-4
27. Phelan J, Coll F, McNerney R et al (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14:31. https://doi.org/10.1186/s12916-016-0575-9
28. Silvino AC, Costa GL, Araujo FC et al (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172

29. White RR, Ponsford AH, Weekes MP et al (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11): e1005977. https://doi.org/10.1371/journal.ppat.1005977
30. Hawkey J, Ascher DB, Judd LM et al (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4(3). https://doi.org/10.1099/mgen.0.000165
31. Holt KE, McAdam P, Thai PVK et al (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50(6):849–856. https://doi.org/10.1038/s41588-018-0117-9
32. Karmakar M, Globan M, Fyfe JAM et al (2018) Analysis of a Novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 198(4):541–544. https://doi.org/10.1164/rccm.201712-2572LE
33. Vedithi SC, Malhotra S, Das M et al (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1
34. Karmakar M, Rodrigues CHM, Holt KE et al (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS One 14(5):e0217169. https://doi.org/10.1371/journal.pone.0217169
35. Ascher DB, Wielens J, Nero TL et al (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
36. Jubb HC, Higueruelo AP, Ochoa-Montano B et al (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
37. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
38. Pandurangan AP, Ochoa-Montano B, Ascher DB et al (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45(W1):W229–W235. https://doi.org/10.1093/nar/gkx439
39. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42(Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
40. Douglas EVP, Carlos HMR, David BA et al (2020) mCSM-membrane: predicting the effects of mutations on transmembrane proteins, Nucleic Acids Research, gkaa416. https://doi.org/10.1093/nar/gkaa416
41. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300
42. Rodrigues CHM, Myung Y, Pires DEV et al (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res 47(W1):W338–W344. https://doi.org/10.1093/nar/gkz383
43. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 44(W1):W469–W473. https://doi.org/10.1093/nar/gkw458
44. Yoochan M, Carlos HMR, David BA, Douglas EVP et al (2020) mCSM-AB2: guiding rational antibody design using graphbased signatures, Bioinformatics. 36(5):1453–1459. https://doi.org/10.1093/bioinformatics/btz779
45. Yoochan M, Douglas EVP, David BA et al. mmCSM-AB: guiding rational antibody engineering through multiple point mutations, Nucleic Acids Research, gkaa389. https://doi.org/10.1093/nar/gkaa389
46. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45(W1):W241–W246. https://doi.org/10.1093/nar/gkx236
47. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
48. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44(W1):W557–W561. https://doi.org/10.1093/nar/gkw390
49. Douglas EVP et al (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC genomics (12) No. S4. BioMed Central
50. Douglas EVP, Raquel CM-M, Carlos HS, Frederico FC, Wagner M Jr (2013) aCSM: noise-free graphbased signatures to large-scale receptor-based ligand prediction, Bioinformatics 29(7):855–861. https://doi.org/10.1093/bioinformatics/btt058

51. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311. https://doi.org/10.1093/nar/29.1.308
52. Stenson PD, Mort M, Ball EV et al (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136(6):665–677. https://doi.org/10.1007/s00439-017-1779-6
53. Landrum MJ, Lee JM, Benson M et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46(D1):D1062–D1067. https://doi.org/10.1093/nar/gkx1153
54. Karczewski KJ, Francioli LC, Tiao G et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv:531210. https://doi.org/10.1101/531210
55. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3): e1001779. https://doi.org/10.1371/journal.pmed.1001779
56. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46(5):2699. https://doi.org/10.1093/nar/gky092
57. Rose PW, Prlic A, Altunkaya A et al (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 45(D1):D271–D281. https://doi.org/10.1093/nar/gkw1000
58. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
59. Witten IH, Frank E, Hall MA et al (2016) Data mining, fourth edition: practical machine learning tools and techniques. Morgan Kaufmann, Burlington

Bibliography

[1] Jin Gao, Wen-Xi Li, Si-Qian Feng, Yun-Sheng Yuan, Da-Fang Wan, Wei Han, and Yan Yu. A protein–protein interaction network of transcription factors acting during liver cell proliferation. *Genomics*, 91(4):347–355, 2008.

[2] Dana Chuderland and Rony Seger. Protein-protein interactions in the regulation of the extracellular signal-regulated kinase. *Molecular biotechnology*, 29(1):57–74, 2005.

[3] Charlotte Nicod, Amir Banaei-Esfahani, and Ben C Collins. Elucidation of host–pathogen protein–protein interactions to uncover mechanisms of host cell rewiring. *Current opinion in microbiology*, 39:7–15, 2017.

[4] Christian M Paumi, Javier Menendez, Anthony Arnoldo, Kim Engels, Kavitha Ravee Iyer, Safia Thaminy, Oleg Georgiev, Yves Barral, Susan Michaelis, and Igor Stagljar. Mapping protein-protein interactions for the yeast abc transporter ycf1p by integrated split-ubiquitin membrane yeast two-hybrid analysis. *Molecular cell*, 26(1):15–25, 2007.

[5] Michael PH Stumpf, Thomas Thorne, Eric De Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[6] Roberto Mosca, Arnaud Céol, and Patrick Aloy. Interactome3d: adding structural details to protein networks. *Nature methods*, 10(1):47, 2013.

[7] Joël Janin, Ranjit P Bahadur, and Pinak Chakrabarti. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2):133, 2008.

[8] Jan Gruber, Alexander Zawaira, Rhodri Saunders, C Paul Barrett, and Martin EM Noble. Computational analyses of the surface properties of protein–protein interfaces. *Acta Crystallographica Section D: Biological Crystallography*, 63(1):50–57, 2007.

[9] Nan Zhao, Bin Pang, Chi-Ren Shyu, and Dmitry Korkin. Charged residues at protein interaction interfaces: unexpected conservation and orchestrated divergence. *Protein Science*, 20(7):1275–1284, 2011.

[10] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[11] Harry Jubb, Tom L Blundell, and David B Ascher. Flexibility and small pockets at protein–protein interfaces: new insights into druggability. *Progress in biophysics and molecular biology*, 119(1):2–9, 2015.

[12] Juan Fernández-Recio. Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):680–698, 2011.

[13] Bingding Huang and Michael Schroeder. Using protein binding site prediction to improve protein docking. *Gene*, 422(1-2):14–21, 2008.

[14] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4):e43, 2007.

[15] George R Bickerton, Alicia P Higueruelo, and Tom L Blundell. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the piccolo database. *BMC bioinformatics*, 12(1):1–15, 2011.

[16] George Richard James Bickerton. *Molecular characterization and evolutionary plasticity of protein-protein interfaces*. PhD thesis, University of Cambridge, 2011.

[17] Adrian M Schreyer and Tom L Blundell. Credo: a structural interactomics database for drug discovery. *Database*, 2013, 2013.

[18] Byungkook Lee and Frederic M Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4, 1971.

[19] Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia, and Michael L Tress. Progress and challenges in predicting protein–protein interaction sites. *Briefings in bioinformatics*, 10(3):233–246, 2009.

[20] Xiong Jiao and Shoba Ranganathan. Prediction of interface residue based on the features of residue interaction network. *Journal of theoretical biology*, 432:49–54, 2017.

[21] Zhe Zhang, Shawn Witham, and Emil Alexov. On the role of electrostatics in protein–protein interactions. *Physical biology*, 8(3):035001, 2011.

[22] Harry C Jubb, Alicia P Higueruelo, Bernardo Ochoa-Montaño, Will R Pitt, David B Ascher, and Tom L Blundell. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *Journal of molecular biology*, 429(3):365–371, 2017.

[23] Gregory A Petsko and John R Yates III. Analyzing molecular interactions. *Current protocols in bioinformatics*, 36(1):8–1, 2011.

[24] Jeremy M Berg, John L Tymoczko, Lubert Stryer, JM Berg, JL Tymoczko, and L Stryer. Biochemistry: International version (hardcover), 2002.

[25] Brian C Cunningham and James A Wells. High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908): 1081–1085, 1989.

[26] Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, 1998.

[27] Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, and David B Ascher. mcsm-ppi2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[28] Antonio Del Sol and Paul O'Meara. Small-world network approach to identify key residues in protein–protein interaction. *Proteins: Structure, Function, and Bioinformatics*, 58(3):672–682, 2005.

[29] Ozlem Keskin, Buyong Ma, and Ruth Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of molecular biology*, 345(5):1281–1294, 2005.

[30] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–2198, 1999.

[31] Yanay Ofran and Burkhard Rost. Protein–protein interaction hotspots carved into sequences. *PLoS Comput Biol*, 3(7):e119, 2007.

[32] Xiang Li, Ozlem Keskin, Buyong Ma, Ruth Nussinov, and Jie Liang. Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *Journal of molecular biology*, 344(3):781–795, 2004.

[33] Deepa Rajamani, Spencer Thiel, Sandor Vajda, and Carlos J Camacho. Anchor residues in protein–protein interactions. *Proceedings of the National Academy of Sciences*, 101(31):11287–11292, 2004.

[34] Alessia David, Rozami Razali, Mark N Wass, and Michael JE Sternberg. Protein–protein interaction sites are hot spots for disease-associated nonsynonymous snps. *Human mutation*, 33(2):359–363, 2012.

[35] Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure*, 23 (7):1362–1369, 2015.

[36] Alexander Gress, V Ramensky, and Olga V Kalinina. Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes. *Oncogenesis*, 6(9):e380, 2017.

[37] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I Karras, Yang Wang, et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660, 2015.

[38] Alessia David and Michael JE Sternberg. The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *Journal of molecular biology*, 427(17):2886–2898, 2015.

[39] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[40] Iain H Moal and Juan Fernández-Recio. Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, 2012.

[41] Sherlyn Jemimah, K Yugandhar, and M Michael Gromiha. Proximate: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, 33(17):2787–2788, 2017.

[42] Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. Ab-bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[43] Quanya Liu, Peng Chen, Bing Wang, Jun Zhang, and Jinyan Li. dbmpikt: a database of kinetic and thermodynamic mutant protein interactions. *Bmc Bioinformatics*, 19(1):1–7, 2018.

[44] Kurt S Thorn and Andrew A Bogan. Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, 2001.

[45] Joan Segura and Narcis Fernandez-Fuentes. Pcrpi-db: a database of computationally annotated hot spots in protein interfaces. *Nucleic acids research*, 39(suppl_1): D755–D760, 2011.

[46] MD Shaji Kumar, K Abdulla Bava, M Michael Gromiha, Ponraj Prabakaran, Koji Kitajima, Hatsuho Uedaira, and Akinori Sarai. Protherm and pronit: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic acids research*, 34(suppl_1):D204–D206, 2006.

[47] Joicymara S Xavier, Thanh-Binh Nguyen, Malancha Karmarkar, Stephanie Portelli, Pâmela M Rezende, João PL Velloso, David B Ascher, and Douglas EV Pires. Thermomutdb: a thermodynamic database for missense mutations. *Nucleic Acids Research*, 49(D1):D475–D479, 2021.

[48] Shuanghong Huo, Irina Massova, and Peter A Kollman. Computational alanine scanning of the 1: 1 human growth hormone–receptor complex. *Journal of computational chemistry*, 23(1):15–27, 2002.

[49] Yosef Y Kuttner and Stanislav Engel. Protein hot spots: the islands of stability. *Journal of molecular biology*, 415(2):419–428, 2012.

[50] Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002.

[51] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.

[52] Solène Grosdidier and Juan Fernández-Recio. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC bioinformatics*, 9 (1):447, 2008.

[53] Steven J Darnell, Laura LeGault, and Julie C Mitchell. Kfc server: interactive forecasting of protein interaction hot spots. *Nucleic acids research*, 36(suppl_2): W265–W269, 2008.

[54] Nurcan Tuncbag, Attila Gursoy, and Ozlem Keskin. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520, 2009.

[55] Douglas EV Pires, David B Ascher, and Tom L Blundell. mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30 (3):335–342, 2013.

[56] Lin Wang, Wenjuan Zhang, Qiang Gao, and Congcong Xiong. Prediction of hot spots in protein interfaces using extreme learning machines with the information of spatial neighbour residues. *IET systems biology*, 8(4):184–190, 2014.

[57] Zhenhua Li, Ying He, Limsoon Wong, and Jinyan Li. Burial level change defines a high energetic relevance for protein binding interfaces. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 12(2):410–421, 2015.

[58] Alexandra Shulman-Peleg, Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC biology*, 5(1):43, 2007.

[59] Catherine L Worth, Robert Preissner, and Tom L Blundell. Sdm—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research*, 39(suppl_2):W215–W222, 2011.

[60] Reed ES Harrison, Rohith R Mohan, Ronald D Gorham Jr, Chris A Kieslich, and Dimitrios Morikis. Aesop: A python library for investigating electrostatics in protein interactions. *Biophysical journal*, 112(9):1761–1766, 2017.

[61] Grant Thiltgen and Richard A Goldstein. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PloS one*, 7(10):e46084, 2012.

[62] Lawrence A Loeb and Fred C Christians. Multiple mutations in human cancers. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 350 (1):279–286, 1996.

[63] Keith R Loeb and Lawrence A Loeb. Significance of multiple mutations in cancer. *Carcinogenesis*, 21(3):379–385, 2000.

[64] Douglas EV Pires, David B Ascher, and Tom L Blundell. mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30 (3):335–342, 2014.

[65] Douglas EV Pires, David B Ascher, and Tom L Blundell. Duet: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research*, 42(W1):W314–W319, 2014.

[66] Arun Prasad Pandurangan, Bernardo Ochoa-Montaño, David B Ascher, and Tom L Blundell. Sdm: a server for predicting effects of mutations on protein stability. *Nucleic acids research*, 45(W1):W229–W235, 2017.

[67] Gen Li, Shailesh Kumar Panday, and Emil Alexov. Saafec-seq: A sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability. *International Journal of Molecular Sciences*, 22(2):606, 2021.

[68] Ivan Getov, Marharyta Petukh, and Emil Alexov. Saafec: predicting the effect of single point mutations on protein folding free energy using a knowledge-modified mm/pbsa approach. *International journal of molecular sciences*, 17(4):512, 2016.

[69] Vincent Frappier, Matthieu Chartier, and Rafael J Najmanovich. Encom server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic acids research*, 43(W1):W395–W400, 2015.

[70] Emidio Capriotti, Piero Fariselli, and Rita Casadio. I-mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research*, 33(suppl_2):W306–W310, 2005.

[71] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

[72] Cunliang Geng, Anna Vangone, Gert E Folkers, Li C Xue, and Alexandre MJJ Bonvin. isee: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics*, 87(2):110–119, 2019.

[73] Yves Dehouck, Jean Marc Kwasigroch, Marianne Rooman, and Dimitri Gilis. Beatmusic: prediction of changes in protein–protein binding affinity on mutations. *Nucleic acids research*, 41(W1):W333–W339, 2013.

[74] Minghui Li, Franco L Simonetti, Alexander Goncearenco, and Anna R Panchenko. Mutabind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic acids research*, 44(W1):W494–W501, 2016.

[75] Samuel Genheden and Ulf Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert opinion on drug discovery*, 10(5):449–461, 2015.

[76] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology*, 429(3):426–434, 2017.

[77] Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R Panchenko, and Minghui Li. Mutabind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience*, 23(3):100939, 2020.

[78] Daniel FAR Dourado and Samuel Coulbourn Flores. A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2681–2690, 2014.

[79] Matteo Tiberti, Alessandro Pandini, Franca Fraternali, and Arianna Fornili. In silico identification of rescue sites by double force scanning. *Bioinformatics*, 34(2): 207–214, 2017.

[80] Harry C Jubb, Arun P Pandurangan, Meghan A Turner, Bernardo Ochoa-Montano, Tom L Blundell, and David B Ascher. Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology*, 128:3–13, 2017.

[81] Alícia P Higueruelo, Adrian Schreyer, G Richard J Bickerton, Will R Pitt, Colin R Groom, and Tom L Blundell. Atomic interactions and profile of small molecules disrupting protein–protein interfaces: the timbal database. *Chemical biology & drug design*, 74(5):457–467, 2009.

[82] Alicia P Higueruelo, Harry Jubb, and Tom L Blundell. Timbal v2: update of a database holding small molecules modulating protein–protein interactions. *Database*, 2013, 2013.

[83] Rachel Torchet, Karen Druart, Luis Checa Ruano, Alexandra Moine-Franel, Hélène Borges, Olivia Doppelt-Azeroual, Bryan Brancotte, Fabien Mareuil, Michael Nilges, Hervé Ménager, et al. The ippi-db initiative: A community-centered database of protein-protein interaction modulators. *Bioinformatics*, 2021.

[84] Avraham Ben-Shimon and Miriam Eisenstein. Computational mapping of anchoring spots on protein surfaces. *Journal of molecular biology*, 402(1):259–277, 2010.

[85] Lidio MC Meireles, Alexander S Dï¿ 1/2mling, and Carlos J Camacho. Anchor: a web server and database for analysis of protein–protein interaction binding pockets for drug discovery. *Nucleic acids research*, 38(suppl_2):W407–W411, 2010.

[86] Eman MM Abdelraheem, Carlos J Camacho, and Alexander Dömling. Focusing on shared subpockets–new developments in fragment-based drug discovery. *Expert opinion on drug discovery*, 10(11):1179–1187, 2015.

[87] Matthew Bartolowits and V Jo Davisson. Considerations of protein subpockets in fragment-based drug design. *Chemical biology & drug design*, 87(1):5–20, 2016.

[88] Benjamin P Cossins and Alastair DG Lawson. Small molecule targeting of protein–protein interactions through allosteric modulation of dynamics. *Molecules*, 20(9): 16435–16445, 2015.

[89] Stephanie Portelli, Moshe Olshansky, Carlos HM Rodrigues, Elston N D'Souza, Yoochan Myung, Michael Silk, Azadeh Alavi, Douglas EV Pires, and David B Ascher. Exploring the structural distribution of genetic variation in sars-cov-2 with the covid-3d online resource. *Nature Genetics*, 52(10):999–1001, 2020.

[90] Houcemeddine Othman, Zied Bouslama, Jean-Tristan Brandenburg, Jorge Da Rocha, Yosr Hamdi, Kais Ghedira, Najet Srairi-Abid, and Scott Hazelhurst. Interaction of the spike protein rbd from sars-cov-2 with ace2: Similarity with sars-cov, hot-spot analysis and effect of the receptor polymorphism. *Biochemical and biophysical research communications*, 527(3):702–708, 2020.

[91] M Shaminur Rahman, M Rafiul Islam, ASM Rubayet Ul Alam, Israt Islam, M Nazmul Hoque, Salma Akter, Md Mizanur Rahaman, Munawar Sultana, and M Anwar Hossain. Evolutionary dynamics of sars-cov-2 nucleocapsid protein and its consequences. *Journal of medical virology*, 93(4):2177–2195, 2021.

[92] Gyanendra Bahadur Chand, Atanu Banerjee, and Gajendra Kumar Azad. Identification of novel mutations in rna-dependent rna polymerases of sars-cov-2 and their implications on its protein structure. *PeerJ*, 8:e9492, 2020.

[93] Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthened sars-cov-2 infectivity. *Journal of molecular biology*, 432(19):5212–5226, 2020.

[94] Malancha Karmakar, Carlos HM Rodrigues, Kathryn E Holt, Sarah J Dunstan, Justin Denholm, and David B Ascher. Empirical ways to identify novel bedaquiline resistance mutations in atpe. *PloS one*, 14(5):e0217169, 2019.

[95] Stephanie Portelli, Jody E Phelan, David B Ascher, Taane G Clark, and Nicholas Furnham. Understanding molecular consequences of putative drug resistant mutations in mycobacterium tuberculosis. *Scientific reports*, 8(1):1–12, 2018.

[96] Sundeep Chaitanya Vedithi, Carlos HM Rodrigues, Stephanie Portelli, Marcin J Skwark, Madhusmita Das, David B Ascher, Tom L Blundell, and Sony Malhotra. Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of rna polymerase in mycobacterium leprae. *Computational and structural biotechnology journal*, 18:271–286, 2020.

[97] Mahdi Ghorbani, Bernard R Brooks, and Jeffery B Klauda. Critical sequence hotspots for binding of novel coronavirus to angiotensin converter enzyme as evaluated by molecular simulations. *The Journal of Physical Chemistry B*, 124(45): 10034–10047, 2020.

[98] Satishkumar Ranganathan Ganakammal, Mahesh Koirala, Bohua Wu, and Emil Alexov. In-silico analysis to identify the role of men1 missense mutations in breast cancer. *Journal of Theoretical and Computational Chemistry*, 19(06):2041002, 2020.

[99] Nagesh Kishan Panchal, Aishwarya Bhale, Vinod Kumar Verma, and Syed Sultan Beevi. Computational and molecular dynamics simulation approach to analyze the impactof xpd gene mutation on protein stability and function. *Molecular Simulation*, 46(15):1200–1219, 2020.

[100] Lisa Jean Ewans, Alison Colley, Carles Gaston-Massuet, Angelica Gualtieri, Mark J Cowley, Mark James McCabe, Deepti Anand, Salil A Lachke, Luigi Scietti, Federico Forneris, et al. Pathogenic variants in plod3 result in a stickler syndrome-like connective tissue disorder with vascular complications. *Journal of medical genetics*, 56(9):629–638, 2019.

[101] Igor Bychkov, Elena Kamenets, Marina Kurkina, Georgiy Rychkov, Alexandra Ilyushkina, Aleksandra Filatova, Darya Guseva, Galina Baydakova, Andrey Nekrasov, Aleksandr Cheblokov, et al. Alkaptonuria in russia: mutational spectrum and novel variants. *European Journal of Medical Genetics*, 64(4):104165, 2021.

[102] G Del Poeta, M Postorino, L Pupo, MI Del Principe, M Dal Bo, T Bittolo, F Buccisano, B Mariotti, E Iannella, L Maurillo, et al. Venetoclax: Bcl-2 inhibition for the treatment of chronic lymphocytic leukemia. *Drugs of Today (Barcelona, Spain: 1998)*, 52(4):249–260, 2016.

[103] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[104] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.

[105] Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[106] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[107] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[108] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[109] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[110] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[111] Martin Riedmiller. Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces*, 16(3):265–278, 1994.

[112] Eibe Frank, Yong Wang, Stuart Inglis, Geoffrey Holmes, and Ian H Witten. Using model trees for classification. *Machine learning*, 32(1):63–76, 1998.

[113] S Kotsiantis, A Kostoulas, S Lykoudis, A Argiriou, and K Menagias. Using data mining techniques for estimating minimum, maximum and average daily temperature values. *International Journal of Mathematical, Physical and Engineering Sciences*, 1(1):16–20, 2008.

[114] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[115] Mohammed J. Zaki and Jr. Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, May 2014. ISBN 9780521766333.

[116] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[117] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[118] David D Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, volume 33, pages 81–93, 1994.

[119] Taqwa Ahmed Alhaj, Maheyzah Md Siraj, Anazida Zainal, Huwaida Tagelsir Elshoush, and Fatin Elhaj. Feature selection using information gain for improved structural-based alert correlation. *PloS one*, 11(11):e0166017, 2016.

[120] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[121] Abhijit Ghatak. Introduction to machine learning. In *Machine Learning with R*, pages 57–78. Springer, 2017.

[122] Bradley Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam, 1982.

[123] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.

[124] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[125] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[126] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[127] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.